Generalized Surface Geometry Estimation in Photometric Stereo and Two-view Stereo Matching

HUNG, Chun Ho

A Thesis Submitted in Partial Fulfilment
of the Requirements for the Degree of
Master of Philosophy
in
Computer Science and Engineering

The Chinese University of Hong Kong
September 2011


THE CHINESE UNIVERSITY OF HONG KONG
23 JAN 2013

Abstract

Digitalizing 3D is of broad interests in computer vision. In this thesis, we study and generalize two traditional problems within this domain - that is, the *Photometric Stereo* and *Stereo Matching* problem.

Traditional photometric stereo has general dark-room requirements in data acquisition, which obstructs its employment in many circumstances that do not allow light manipulation. We present a new method to mitigate this restriction with a handy configuration, which substitutes the traditional chrome/diffuse spheres with mirror spheres to capture environment lights. Instead of physically controlling light, our method simply captures the complex environment and utilizes it in photometric stereo. This strategy was found very effective to accomplish surface reconstruction in many scenes that are traditionally regarded as impossible to utilize. It greatly generalizes photometric stereo.

We also propose a robust depth and motion estimation method taking the input of a binocular video. We focus on solving the challenging motion-depth boundary consistency preservation problem, making the estimates in a long sequence not quickly diverge. Solving this problem actually involves tackling respectively a number of technical issues, including the connection establishment between motion and depth, general optimization framework definition, structure consistency preservation in multiple frames, and global-

constraint employment for error correction. We address these issues in a unified stereo-motion estimation framework. Our main contributions include developing motion trajectories, which robustly link multiple frames for prior establishment, building novel edge occurrence maps, and incorporating the anisotropic smoothing priors into the final objective functions to regularize optimization.

摘要

三維數字化是計算機視覺研究之中的一個重要領域。在這篇論文中，我們研究和廣義化在這傳統領域內的兩個問題——*光度立體視覺* 和 *立體匹配* 問題。

傳統的光度立體技術對於數據採集具有普遍黑房要求，從而妨礙其在很多下不允許光操縱的情況。我們提出了一種新的方法來減輕這種限制的一個方便的配置，以鏡球替代傳統的鉻/擴散球來捕捉環境光。而不是的控制光，我們的方法能簡單地捉取複雜的環境，並利用它進行光度立體技術。這一策略被發現能非常有效的完成表面重建，亦包括許多在傳統上被認爲不可能的場景。它極大地概括光度立體。

我們還提出了一個採取雙目視頻輸入進行穩固的深度和運動估計方法。我們專注於解決具挑戰性的運動及深度邊界的一致性維護問題，使其在很長的視頻序列中不會迅速發散。解決這個問題實際上涉及到數個技術問題，包括建立深度與運動之間的關係、定義一般優化框架、保存在多幀上的結構一致性，以及全局約束糾錯。爲解決這些問題，我們建立一個統一的立體/運動估計框架。我們的主要貢獻包括發展運動軌跡，這能健壯地鏈接多個幀以建立先驗，建設新的邊緣發生圖，並結合各向異性平滑先驗到最終目標功能規範優化。

Thesis/Assessment Committee

Professor WONG, Kin Hong (Chair)

Professor JIA, Jiaya (Thesis Supervisor)

Professor ZHANG, Shengyu (Committee Member)

Doctor Yasuyuki Matsushita (External Examiner)

Acknowledge

I would like to thank Professor JIA Jiaya, who have patiently motivated me and provided thoughtful comments to develop the main idea of the thesis. This thesis would not have been possible without his abundant guidance and invaluable assistance.

Also, my gratitude is devoted to WU Tai-pang and XU Li, without whose knowledge and assistance this study would not have been successful.

I thank my fellow labmates in our group: YAN Qiong, LU Cewu, XU Yi, DAI Zhenlong, QIN Zenglu and WANG Liwei for the stimulating discussions, and for the sleepless nights spent working together to meet deadlines.

Last but not least, I would like to thank my family. Their constant inspiration and support kept me focused and walk through frustration.

Finally, I wish to thank everyone who directly or indirectly offered his or her help to this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Three dimensional data from images is of central importance in the field of computer vision. Its applications include virtual reality and heritage recording. In this thesis, we focus on generalizing two classes of these problems - *photometric stereo* and *stereo matching*.

Photometric stereo (PS) is a method for shape estimation using several intensity images obtained under different lighting conditions. Fig. 1.1 shows an example set of input images. The set of images are also required to be pixel-aligned, so the pixel correspondences from multiple images can be used to infer the geometry. The main light source(s) in each image, which is usually a distant point light source, has to be calibrated. Traditional PS computes per-pixel surface normal and reflectance (or texture) by solving a series of shading equations using the measured pixel values [40]. By integrating the estimated normal map, one can obtain a per-pixel dense depth map, and hence the 3D of the captured scene.

While PS is known to be capable of accurately recovering surface details (or the high-frequency components), in most configurations the computed 3D data suffer from a low-frequency distortion, or are "flattened"

Figure 1.1: 5 of the input images of the vase example.

along the view direction (Fig. 1.2). This is often due to the violation of the dark-room assumption in data acquisition- that is, there is no prominent additional lighting besides the main light sources. Such a violation mainly originates from the presence of extra complex lighting, ambient lighting, and background reflection. Therefore, photo shooting needs to be restricted in a human-configured dark-room-like environment, where the aforementioned extra lighting does not exist. This inconvenience actually impairs the usefulness and popularity of PS.

We in this thesis introduce the use of a *mirror sphere* to generalize PS, in a sense that photo taking can now be in *any* environment. The principle is that, instead of identifying the main light sources, we consider the whole environment which is directly *captured* by the mirror sphere. As shown in Fig. 1.2 and later in Chapter 2.8, its usability can effectively resolve the aforementioned problems in the traditional PS settings. A more general shading model is derived, and dense normal maps can be efficiently computed using our novel discrete-continuous optimization framework (Chapter 2.4.1).

Stereo matching aims to find corresponding pixels between two (or more) images, *i.e.*, pixels projected from the same 3D point into the images. In the binocular case, *image rectification* is often operated on the images so as to

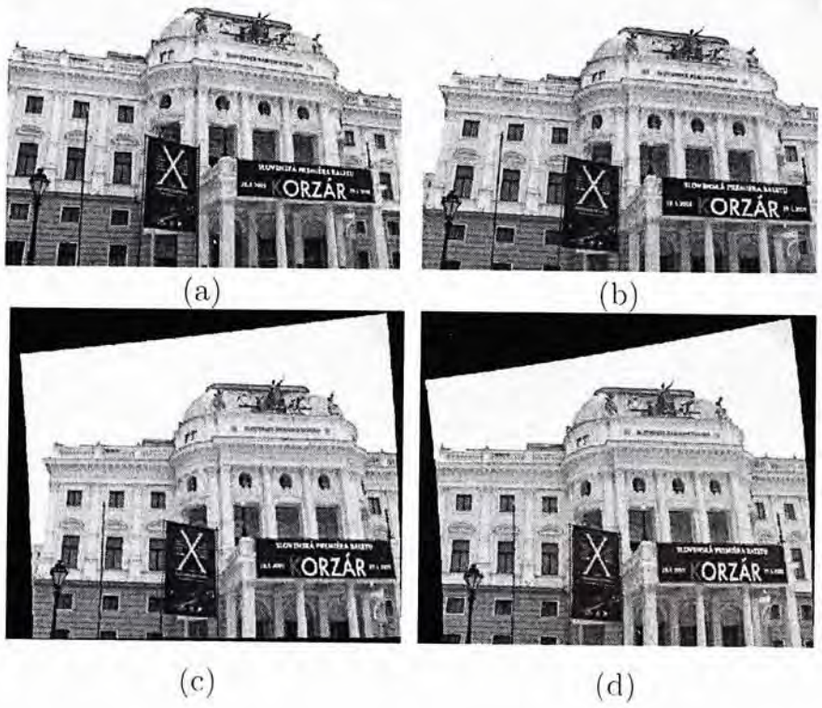
(a) (b) (c) (d)

Figure 1.2: (a) and (b) show the computed surface using traditional PS method and our method respectively. (c) and (d) are another view of (a) and (b) respectively. The surfaces displayed in (a) and (c) are clearly "flattened".

reduce the correspondence problem to a 1D search (Fig. 1.3). The principle is that, any image pair can be transformed to a "parallel camera geometry" such that each 3D point in the scene will lie on the same horizontal line in the two images. Based on the computed disparity, the depth of each pixel can be determined.

Most stereo algorithms can be roughly classified into two categories: *local* and *global* methods. Local methods compute each pixel's disparity independently over a support region. The matching costs are aggregated over the region, and the disparity level with the minimal cost is selected as the output of the pixel [1]. Global methods formulate a global energy function which consists of constraints such as color similarity and disparity smoothness. The function is minimized with dynamic programming [6], graph cuts [7] or belief propagation [15].

While the global optimization approach can obtain fairly spatially smooth disparity maps, foreground fattening in ambiguous region is a major problem.



(a) (b)

(c) (d)

Figure 1.3: (a) and (b) are a left/right image pair. (c) and (d) show the image rectification results.

(a) (b) (c)

Figure 1.4: (a) and (b) show the input left and right images respectively. (c) shows the computed disparity using a global optimization framework. The pixel value is proportional to the computed disparity value.

(a) (b) (c)

Figure 1.5: (a)-(c) show 3 consecutive frames of depths using global optimization method for each pair of left/right frames individually. Obvious flickering artifacts present.

As shown in Fig. 1.4 (c), it is difficult for the global optimizer to determine whether the pixels on the textureless wall should be on the foreground or the background. To compute depths of a binocular video, a straightforward approach is to compute disparity map for each pair of left/right frame. However, the aforementioned fattening problem will result in undesirable flickering artifacts (Fig. 1.5).

We propose the *trajectory-based structure profile* to address the problem. First, we compute the motion trajectory which considers occlusion to identify multi-frame pixel correspondences. An edge occurrence map is then computed for each frame by simultaneously considering multi-frame edges. By incorporating multi-frame information, salient and consistent object boundary is preserved, while occasional errors and redundant edges are suppressed. By minimizing a complex global objective using an iterative multi-grid framework, we obtain temporally consistent depth map which can well preserve discontinuity across object boundary when necessary.

□ **End of chapter.**

Chapter 2

Generalized Photometric Stereo

2.1 Problem Description

Digitalizing 3D objects from still images is one of the core problems of computer vision. To recover a highly detailed surface, photometric stereo (PS) is known to be one of the most effective methods. Since Woodham [40] introduced the concept of photometric stereo, plenty of research work was conducted to improve the estimation quality, computation robustness, and the reflectance models.

One critical problem that hinders PS from being widely employed is the stringent lighting condition requirement. Generally, the number of light sources as well as the corresponding lighting direction have to be known in order for PS to work. Therefore, data acquisition has to be performed inside a dark-room or in an enclosed workspace where complex lighting components, such as background inter-reflection, do not exist. This is an inherent problem that most of the current PS systems are facing.

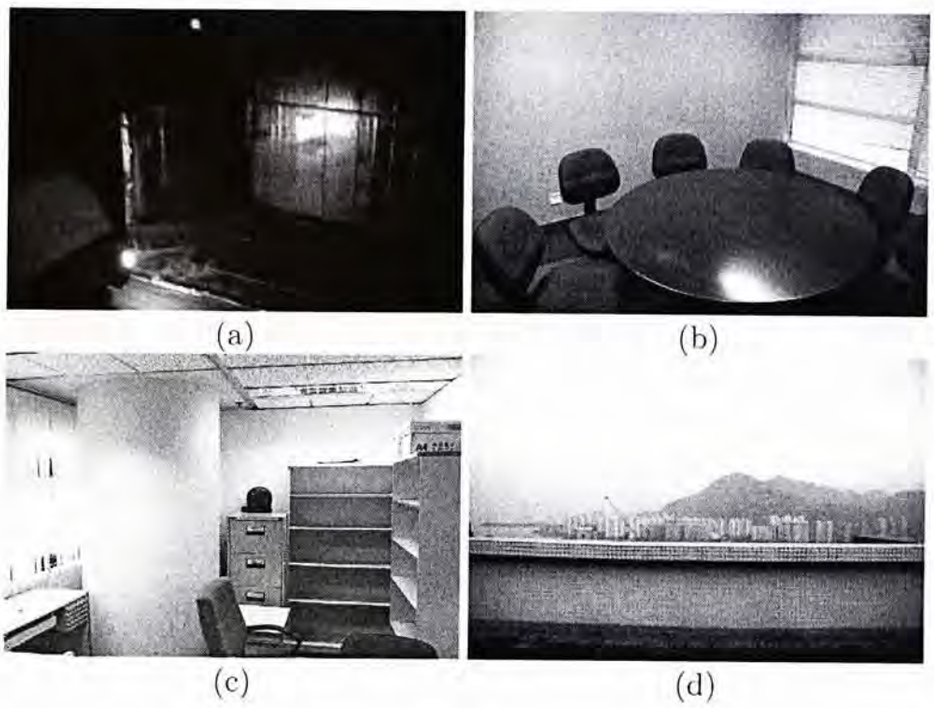


Figure 2.1: Typical scenes that make photometric stereo estimation difficult.

As a result, when unexpected light is emitted or reflected in the environment, as exemplified in Fig. 2.1, PS may fail. Fig. 2.1 contains a laboratory in (a) where all lights are deliberately switched off but there is still ambient light reflected from wall and furniture when turning on a spotlight. Two indoor scenes in (b)-(c) contain different levels of inter-reflection. Light passing through windows further complicates it. The last example shown in (d) is a ubiquitous open area in which it is more difficult to control lighting.

These examples show that in general complex lighting is hard to avoid. To make PS more general, instead of physically controlling the environment, which sometimes is impossible, a better way is to naturally include the captured light in estimation such that ubiquitous ambient light will not be decisively harmful, but on the contrary help produce decent results.

Following this line, in this thesis, we propose a simple but very effective technique to handle complex environment lighting in PS, bringing the image

capturing process out of human-set laboratories. Our method does not assume that only sun light [32] exists, but instead uses a low-cost mirror sphere configuration to accurately and quickly *capture* more complex environment lighting. Our main contribution is the employment of a new setup in PS to capture light and sample it for normal and surface estimation.

Characteristics of our method Except for photometric still images, our system has the following main advantages compared to traditional systems.

(a) Traditional methods need to detect the main direct light sources while our method does not. We simply capture the whole environment lighting map for each image without finding and distinguishing major light sources.

(b) Traditional methods generally move or vary the configuration of major light sources to produce photometric parallax. In our method, any alteration in the environment, which is not restricted to major light sources, is usable to produce photometric parallax.

Assumptions The assumptions that apply to our work are those also used in traditional PS.

(i) The class of objects that we handle depends on the descriptiveness of the chosen analytical model. The Lambertian model is discussed in detail in this thesis.

(ii) All environment light sources are distant.

These conditions are not strict. For (i), we found that the Lambertian model in the environment lighting configuration can satisfyingly handle many objects with our outlier-rejection framework, as shown in this thesis. For (ii), empirically we produce decent results inside a small room where lights are

put close to the objects. Both of these requirements are no more in terms of both quantity and degree than those used in traditional PS.

2.2 Related Work

In this section, we first briefly review photometric stereo methods. As our focus in this thesis is to propose a new setting for general light capturing, we will discuss previous light inference/estimation steps.

Classical PS Methods Given a static Lambertian scene as well as the number of distant light sources and the corresponding lighting directions, Woodham [40] showed that at least three images are required to recover the surface orientation for each pixel. In [12], four images were used where shadow and specularity are identified as outliers. To handle more objects, the Torrance-Sparrow model [35] was proposed to deal with non-Lambertian surfaces. With the reflectance information, a hybrid reflectance model was introduced in [26] to extract local surface orientation along with reflectance.

Modern PS Methods Compared with these classical methods, modern PS techniques grant a higher degree of freedom for both the surface/reflectance representation and the experimental settings.

Representative work includes the example based methods [18, 16] to transfer normals from reference or virtual objects to the target one; the glossy surface method using cast shadow for normal estimation [11]; bivariate approximation of the isotropic reflectance function to model spatially varying reflectance [3]; the sub pixel surface normal estimation method [36] using Expectation Maximization (EM); the method to tackle the anisotropic reflectance effect [20]; the dense frame robust photometric stereo method [41] for casual data acquisition; the colored lighting method [17] for dynamic ob-

jects; the consensus method [19] to get rid of analytical models; and the self calibration method [34] to automatically calibrate lighting information and scene radiance simulataneously.

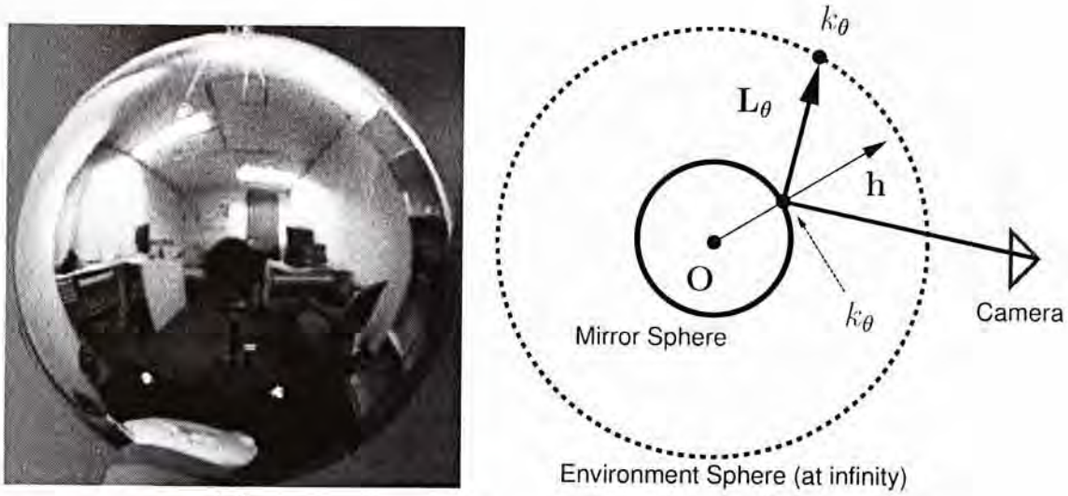
Recently, multiview geometry was incorporated to avail constructing wider-view 3D surfaces [24]. In [27], Nehab *et al.* proposed Multiview photometric stereo (MVPS), and used a triangulation scanner and a photometric stereo scanner to capture 3D models. Esteban *et al.* [14] combined MVS and PS to recover complete surfaces with details.

Light Calibration using Spheres The above methods either require light calibration using chrome spheres or estimate unknown lighting in a controlled environment. Many of them assume that there is a countable number of *direct* light sources. As aforementioned, this is one of the inherent problems of PS, which impairs the usefulness and population of this technique.

Wu *et al.* [41] estimated lighting by identifying saturated pixels on a mirror sphere. With an additional diffuse sphere, Goldman *et al.* [16] computed relative light intensity. Ubiquitous indoor or outdoor environments, where the single point-source assumption does not hold, do not satisfy their assumptions.

There are a few methods, such as the ones described in [32, 33], to preliminarily study PS for outdoor scenes. These approaches address some specific problems. In contrast, we propose a simple but very general technique to allow data acquisition in a primarily uncontrolled environment. Lights coming from almost all possible directions that are visible to the camera are considered.

Our work is essentially different from the method of [5]. In [5], no light calibration device is employed and unknown lighting and object shape are algorithmically *estimated*. Due to the high difficulty of the problem and



(a) (b)

Figure 2.2: A mirror sphere for capturing environment lighting. (b) shows the relationship between the mirror sphere and the environment sphere, where **O** is the *ideal* location where the object is placed, **h** is the angle bisector between the lighting direction $\mathbf{L}_\theta$ and the viewing direction. k_θ is the incident intensity along $\mathbf{L}_\theta$.

the ambiguity of solutions, it has to assume a smooth reflectance function that can be approximated by second-order spherical harmonics and require user input to handle the generalized bas-relief (GBR) ambiguity manually. Computation time can also be an issue. In comparison, we propose a handy, simple, and general method with mirror spheres to directly *capture* lights without assuming particular environment types. The photometric stereo estimation is *automatic* and can produce high quality results in a short time.

2.3 Photometric Stereo with Environment Lighting

A mirror sphere was used in computer graphics for environment capturing and photo-realistic rendering [13]. One example is shown in Fig. 2.2(a). Its

usability in PS however was not explored. We show that its proper employment in PS can enable the capture of incoming direct (*e.g.*, moving handheld spotlight, fluorescent tube) and indirect (*e.g.*, inter-reflections) illumination, thus greatly benefiting PS in general scenes.

The mirror sphere shown in Fig. 2.2(a) captures complex lights, even from the emissive black bodies and the environment, in addition to the major light sources that were considered in conventional PS. The rich light information allows alleviating the restriction of only detecting major lights through dense color sampling from the sphere. For the demonstrated difficult environments, all these samples can be regarded as light sources, which make the following estimation process very accurate.

Fig. 2.2(b) illustrates the relationship between lighting direction $\mathbf{L}_\theta \in \mathbb{R}^3$ and the corresponding intensity $k_\theta \in \mathbb{R}$. It indicates that performing light capturing using a mirror sphere does not need to distinguish actual *direct* and *indirect* lights. They are uniformly taken into consideration in our method. We show in Chapter 2.8 that with this strategy we can produce decent normal maps even if the lighting condition is unknown or uncontrollable.

Similar to the notations used in [28], the intensity I observed at a surface point can be modeled as an integration

$$I = \int_{\theta \in \Omega} f(\mathbf{L}_\theta, \mathbf{V}) k_\theta \mathbf{N}^T \mathbf{L}_\theta d\theta, \tag{2.1}$$

where $f(\cdot)$ is the BRDF at the surface point, Ω is a space containing all possible orientations, $\mathbf{L}_\theta$ is a lighting direction defined by orientation θ, k_θ is the incident intensity at θ, $\mathbf{V}$ is the viewing direction at the surface point, and $\mathbf{N}$ is the corresponding surface normal.

2.4 Estimating Surface Normals

For simplicity's sake, we deal with the widely employed Lambertian model, and show that decent results can be produced in a large variety of uncontrolled scenes, which is difficult, if not impossible, to accomplish in prior work. Eq. (2.1) is expressed as

$$I = \sum_{\theta \in \Omega'} \rho k_\theta \mathbf{N}^T \mathbf{L}_\theta, \tag{2.2}$$

where ρ is the surface albedo and Ω' is the discrete version of Ω. The discretized Ω' corresponds to the sampled directions on the mirror sphere. From the camera's point of view, I is the illuminance observed at a pixel.

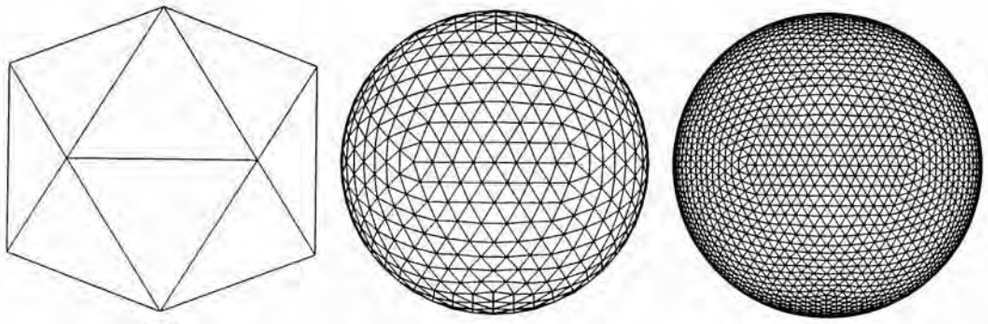
Note that Eq. (2.2) is different from the one used in [33]. We present an explicit model characterized by surface normals and lighting directions, while the method of Shen and Tan [33] is derived based on cosine kernels and coefficients in spherical harmonics to represent data obtained from internet, where recovering dense and accurate normal maps is not the main concern.

With Eq. (2.2), our target is to estimate ρ and $\mathbf{N}$ from a set of captured images I_i, indexed by $i = 1 \cdots n \mid n \geq 3$ with a fixed viewing direction. As $\mathbf{N}^T \mathbf{L}_{\theta_i}$ cannot be negative, Eq. (2.2) can be more accurately expressed as

$$I_i = \sum_{\theta \in \Omega'} \rho k_{\theta_i} \max(\mathbf{N}^T \mathbf{L}_{\theta_i}, 0), \tag{2.3}$$

where k_{θ_i} and $\mathbf{L}_{\theta_i}$ are respectively the incident intensity and the lighting direction defined by θ for image i. The $\max(\cdot)$ operator is used to reject negative energies. Eq. (2.3) describes a single-pixel color formation.

Because of the presence of $\max(\cdot)$, Eq. (2.3) is not continuous in its form. We thus cannot simply apply existing continuous optimization methods, such as gradient descent, to solve for ρ and $\mathbf{N}$. Also, ignoring the $\max(\cdot)$ operator to approximate a linear system can make the negative $\mathbf{N}^T \mathbf{L}_{\theta_i}$ significantly



(a) (b) (c)

Figure 2.3: Icosahedron and its subdivision. (a) A 20-face polyhedron, where the vertices are evenly distributed on a 3D unit sphere that encloses and touches the polyhedron. (b) A 3-time subdivided icosahedron. (c) A 4-time subdivided icosahedron. Each subdivision is done by splitting each face into four equilateral triangles followed by reprojecting the vertices onto a unit 3D sphere.

affect the result. These facts feature in our PS representation. Note in traditional configurations, with only a few light sources, rejecting negative $\mathbf{N}^T\mathbf{L}_{\theta_i}$ is easy by removing dark regions or shadow, while in our problem, the number of negative terms is much larger. We will show later the result comparison with and without removing the negative terms (Fig. 2.8(d) and (c)). In what immediately follows, we propose a two-step approach as the new solver.

2.4.1 Surface Normal and Albedo Estimation

We propose a novel two-step method. In the first step, we estimate an initial normal and a surface albedo by selecting the most suitable candidate from a discrete set of surface normals. Afterwards, an energy function associated with Eq. (2.3) is formulated and is solved based on the initial normal. This strategy can quickly find the appropriate normal estimates for each pixel.

Step 1: Discrete Approximation Fig. 2.3(a) shows an image of an icosahedron. One interesting property of an icosahedron is that the respective vertices are evenly distributed over a unit 3D sphere. This property also applies to differently subdivided versions of the icosahedron (Fig. 2.3(b)-(c)). We compute a set of possible surface orientations that are evenly distributed in all directions and denote them as the candidate normals $\mathcal{N}$. In experiments, we subdivide the icosahedron four times, which yields 2562 different normal candidates. Only half of them (*i.e.*, 1281 normals) are visible to the camera.

For each $\tilde{\mathbf{N}} \in \mathcal{N}$, we solve for ρ by minimizing the following energy

$$E(\rho) = \sum_{i} ||I_i - \sum_{\theta \in \Psi} \rho k_{\theta_i} \tilde{\mathbf{N}}^T \mathbf{L}_{\theta_i}||^2, \tag{2.4}$$

where $\Psi \in \Omega'$ is a space containing all lighting directions making $\tilde{\mathbf{N}}^T \mathbf{L}_{\theta_i} \geq 0$. Since the surface normal is fixed for each $\tilde{\mathbf{N}}$, we can identify Ψ easily. $\max(\cdot)$ disappears because of the introduction of Ψ. Eq. (2.4) can be solved quickly by a least-square method.

After getting all possible $\tilde{\mathbf{N}}$, we pair each $\tilde{\mathbf{N}}$ with the associated albedo $\tilde{\rho}$ and error $E(\tilde{\rho})$. The pair $(\tilde{\mathbf{N}}^0, \tilde{\rho}^0)$ that produces the smallest error is considered the appropriate candidate, which is the input to step 2 as the initial guess.

Provided with a number of images captured in a light-varying environment, we find that a unique global minimum of E always exists for all examples. Since the discrete space Ψ varies w.r.t. $\tilde{\mathbf{N}}$, we show two examples in Fig. 2.4 with plotted errors. They demonstrate that global minima exist and the respective error surfaces are smooth.

Step 2: Continuous Refinement Since the normal space described by $\mathcal{N}$ is discretized by a subdivided icosahedron, after obtaining the initial $\tilde{\mathbf{N}}^0$ and $\tilde{\rho}^0$, we further refine them to increase the pixel-wise estimation accuracy. This

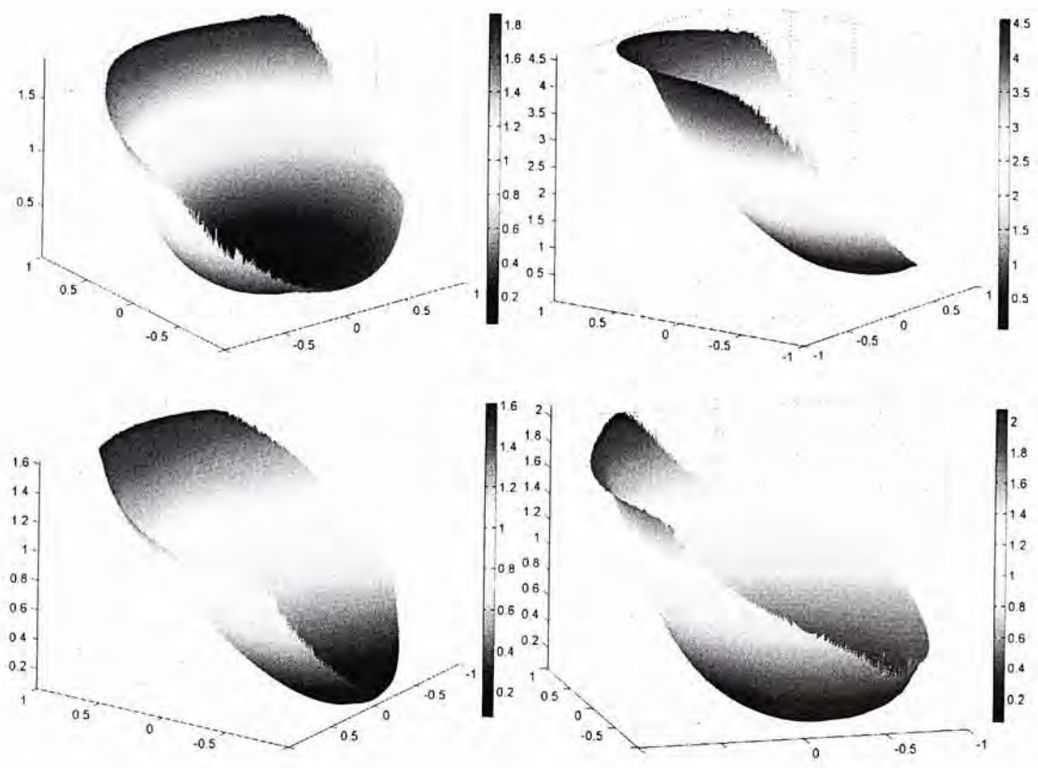


Figure 2.4: Each pair ($\tilde{\mathbf{N}}$, $\tilde{\rho}$) yields an error $E(\tilde{\rho})$. Its plots for two randomly selected pixels are shown, where the z-axis is $E(\tilde{\rho})$ and the x- and y-axes represent the slant and tilt of $\tilde{\mathbf{N}}$. Global minima exist for all test cases, which mean that only one pair of $\tilde{\mathbf{N}}$ and $\tilde{\rho}$ can produce the minimum error, which is reliably selected as the initial guess to step 2.

is done by minimizing Eq. (2.4) using the continuous Levenberg–Marquardt algorithm [30] (one type of gradient decent) starting from $\tilde{\mathbf{N}}^0$ and $\tilde{\rho}^0$. In each iteration, when $\tilde{\mathbf{N}}$ and $\tilde{\rho}$ are updated, the space of Ψ has to be updated too to avoid a negative energy introduced by negative $\tilde{\mathbf{N}}^T\mathbf{L}_{\theta_i}$. However, in practice, we found that the result is almost the same even with fixed Ψ. refinement step.

Coarse-to-Fine Estimation For a single pixel, Eq. (2.4) originally has to be computed $|\mathcal{N}|$ times with $|\Psi|$ possible lighting directions in order to examine the possible pairs of $\tilde{\mathbf{N}}$ and $\tilde{\rho}$, which is computationally expensive. To speed it up, we apply a coarse-to-fine searching algorithm where estimation of $\tilde{\mathbf{N}}^0$ starts from the original icosahedron that contains the smallest number of vertices. After the best candidate normal is found in this level (by applying step 1), we move to a finer level by subdividing the icosahedron and take the current normal estimate (represented by an icosahedron vertex) as the starting position, which is already good except that it lacks a higher degree of accuracy. We search neighboring vertices, which correspond to more normal samples, and move to the one that produces the minimum error in Eq. (2.4). The search is achieved by following the path minimizing Eq. (2.4). When the moving process stops, we go to a finer level. This repeats until the desired subdivision level, which corresponds to the degree of accuracy for the normal estimation, is reached. The process is illustrated in Fig. 2.5.

2.5 Data Acquisition Configuration

We establish the correspondence between our environment lighting calibration system with those that are traditionally employed in a dark-room where only countable major distant point light sources exist. Our major finding is

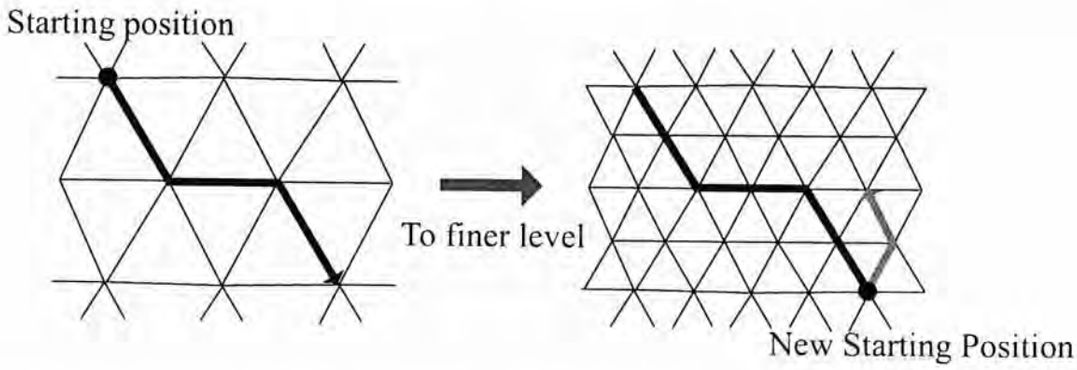


Figure 2.5: Coarse-to-fine estimation. Here shows the subdivided icosahedrons in two levels. With the normal estimate in the coarse level, we continue the search to increase the accuracy in the finer level, which forms a path shown in the left figure. When it stops, the estimate is further passed to the next level shown on the right until the maximum level is reached.

that even with complex environment lighting, if it is faithfully captured, we can find a single virtual light source in an imaginary dark-room that *corresponds* to it. Therefore, the traditional chrome sphere light calibration can be regarded as a *special case* of our configuration.

To prove it, we express Eq. (2.2) as

$$I = \rho \mathbf{N}^T \sum_{\theta \in \Psi} k_\theta \mathbf{L}_\theta = \rho \mathbf{N}^T (\gamma \hat{\mathbf{L}}), \tag{2.5}$$

where $\gamma = || \sum_{\theta \in \Psi} k_\theta \mathbf{L}_\theta ||$ and $\hat{\mathbf{L}} = \frac{1}{\gamma} \sum_{\theta \in \Psi} k_\theta \mathbf{L}_\theta$. γ and $\hat{\mathbf{L}}$ are respectively the *virtual light intensity* and *direction*. With this simplified form, it is possible to solve for $\mathbf{N}$ given three different pairs of γ and $\hat{\mathbf{L}}$. Fig. 2.6 shows the relationship between the image of a mirror sphere and the corresponding virtual lighting in a dark-room.

This observation further implies that we can have much more ways to change lights for photometric stereo in general scenes compared to traditionally only moving the major light sources. We list a few new and intriguing choices as follows, which greatly expand the variety of scenarios to apply PS

and enable new strategies for further exploration.

(1) We can simply alter part of the environment. This can be achieved by powering on/off some ceiling or floor lights in a room, moving the whole system to another place or with a certain angle, or by simply blocking the incident environment light using a black board.

(2) We allow the use of a handheld spotlight even with complex environment lighting. Moving the spotlight can similarly produce photometric parallax.

Fig. 2.6(a) shows an indoor environment allowing lights to be turned on/off. We have also built a trolley that can easily transport the camera, sphere, and the object to different places for photometric stereo without requiring any human controlled light. The second acquisition method was exemplified in Fig. 2.6(c), using a handheld spotlight.

In theory, three images are enough for solving Eq. (2.4) with photometric parallax. However more input generally can help increase accuracy and robustness. Seven or more images are used empirically when noticeable cast shadow arises. The cast shadow issue will be addressed in Chapter 2.7.

2.6 Issues

We discuss possible issues related to the use of mirror spheres for environment light calibration here.

Saturated Pixels The mirror image can be saturated for some pixels when highlights exist. This issue can be readily addressed using the high-dynamic range (HDR) imaging technique [13]. It linearizes the camera response curve and makes the pixel intensity correctly proportional to the ground truth radiance.

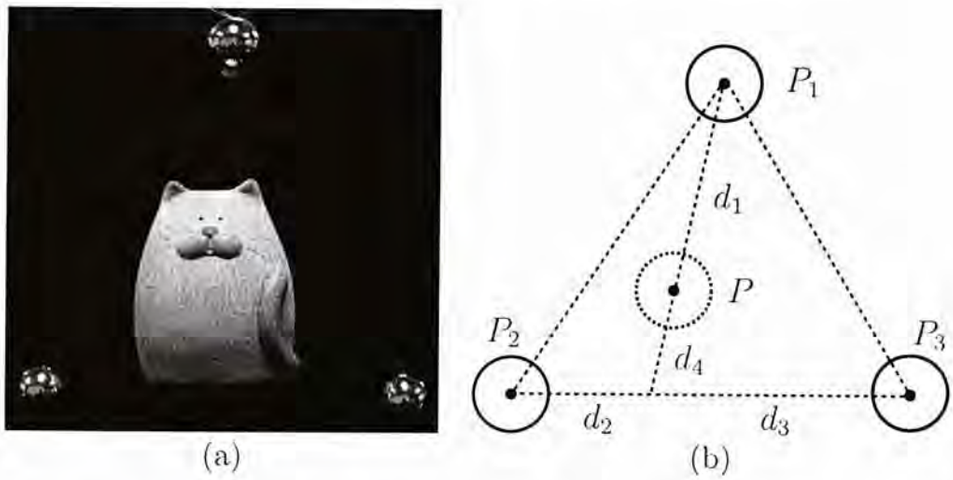
(a) (b) (c) (d)

Figure 2.6: Environment lighting captured by a mirror sphere can be converted into a single virtual light source by Eq. (2.5). (a) and (c) show two mirror spheres. (b) and (d) show the virtual light sources corresponding to them in an imaginary dark-room.

Sampling of Ω' The image of a mirror sphere is the 2D projection of a 3D sphere. It spatially suffers from non-uniform sampling of radiance – that is, the center of the sphere image has a higher sampling rate than the boundary. So we cannot use pixel color directly from the image. Instead, the subdivided icosahedrons are employed, which are aligned with the mirror sphere such that the pixels that coincide with respective vertices are considered for sampling. We subdivide the icosahedron three times, which yields 642 sampling points.

With the regular geometry of a sphere, given a sample location, we can determine its surface normal and take it as the angle bisector to estimate the corresponding $\mathbf{L}_\theta$ using the law of mirror reflection. Before sampling, the mirror sphere image is low-passed by a Gaussian filter to avoid infamous aliasing where the standard deviation is set proportional to the edge length of the subdivided icosahedron. This process can largely reduce the sampling problem.

Placement of Mirror Spheres In practice, the object to be captured cannot be at the same position as the mirror sphere. So, similar to all



(a) (b)

Figure 2.7: (a) The placement of the mirror spheres. (b) An illustration of mirror image interpolation.

other light calibration configurations for PS using chrome/diffuse spheres, estimation errors can be caused by the bias of lighting sampled from the sphere. We propose capturing redundant information from multiple mirror spheres, which are placed at different locations near the object, to alleviate the problem.

We use three mirror spheres on a plane approximately parallel to the image plane. One sphere is above the object while the other two are respectively at the lower left and right corners. One example is shown in Fig. 2.7(a). We blend the three mirror sphere images using triangular interpolation (as illustrated in Fig. 2.7(b)):

$$P = \frac{d_4}{d_1 + d_4} P_1 + \frac{d_1}{d_1 + d_4} \left(\frac{d_3}{d_2 + d_3} P_2 + \frac{d_2}{d_2 + d_3} P_3 \right),$$

where P is the resultant mirror sphere image corresponding to a virtual sphere, P_1-P_3 are the three mirror sphere images, and d_1-d_4 are lengths illustrated in Fig. 2.7(b), coarsely measuring the distances to the center of the object.

(a) (b) (c) (d)

Figure 2.8: Shadow and highlight outliers. One out of 12 input images is shown in (a). The surface results without and with outlier removal are shown in (b) and (d) respectively. If negative $\mathbf{N}^T\mathbf{L}_{\theta i}$ terms are included in computation (described in Chapter 2.4), even with outlier rejection, the result is not good, as shown in (c).

2.7 Outlier Removal

With a non-convex shape or non-Lambertian materials of the captured object, cast shadow and highlight are sometimes unavoidable as shown in Fig. 2.8(a), which could wreck the geometry estimation. For one pixel, if we can reliably find more than three input images, in which the pixel is not in shadow or highlight, the normal and albedo estimates ($\tilde{\mathbf{N}}^0$ and $\tilde{\rho}^0$) can be accurate. To identify a few of these frames, we propose a RANSAC scheme described below (Note that we are now free from negative $\tilde{\mathbf{N}}_0^T\mathbf{L}_{\theta i}$ terms and so it is feasible to apply RANSAC scheme):

1. Randomly choose 3 images from the input and compute a normal-albedo pair $(\tilde{\mathbf{N}}^0, \tilde{\rho}^0)$, using the algorithm described in Chapter 2.4.

2. For each image i, compute the squared error E_i' with respect to the analytical model $E_i' = ||I_i - \sum_{\theta\in\Psi} \tilde{\rho}_0 k_{\theta i} \tilde{\mathbf{N}}_0^T \mathbf{L}_{\theta i}||^2$.

3. Compute the set of inlier images : $T = \{i \mid E_i' \leq \epsilon_i\}$.

4. Re-estimate the normal-albedo pair $(\tilde{\mathbf{N}}_T, \tilde{\rho}_T)$ using all images in T by the algorithm described in Chapter 2.4.

5. Perform step 2 again after replacing $(\tilde{\mathbf{N}}_0, \tilde{\rho}_0)$ by $(\tilde{\mathbf{N}}_T, \tilde{\rho}_T)$. Also update the set of inliers in step 3.

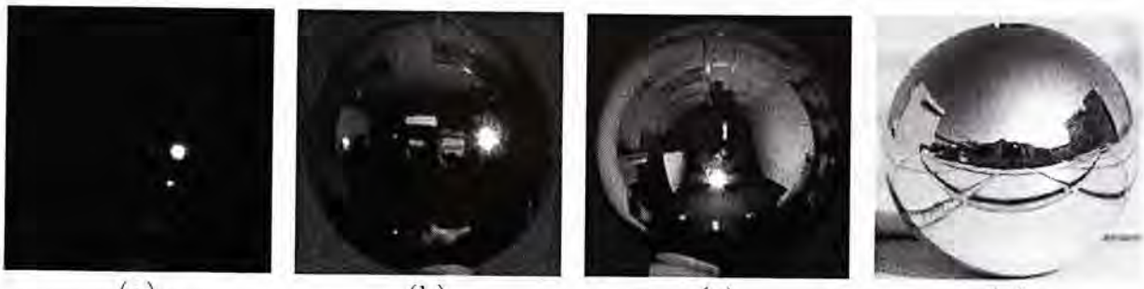
6. Repeat steps 1-5 K times. Keep track of $(\tilde{\mathbf{N}}_T, \tilde{\rho}_T)$ that has the largest set of inliers.

This procedure samples different image triples to identify outliers. The intensity equations are accurately explained by the normal and albedo reported as inliers, given that in the input images highlight and shadow pixels are not the majority. The final per-pixel estimates are chosen as those that correspond to the largest-size inlier set based on a credible error measure. This scheme is found to be very effective to remove problematic pixels in surface estimation.

To deal with brightness variation, the threshold for each image ϵ_i is set to be proportional to the average of the environment light intensity. Empirically, if the number of input images is small, we can replace the random process by exhaustion of all image triples. Fig. 2.8(c) shows that our outlier rejection notably improves the surface estimates.

2.8 Experimental Results

We evaluate our technique in a variety of challenging environments, in which traditional photometric stereo is hard to apply. Fig. 2.9 shows the captured mirror images. The scenes contain different levels of inter-reflection and lighting. Table 2.1 lists the running time. The color coded normal is rendered



(a) (b) (c) (d)

Figure 2.9: Environments for data capturing in our experiments. (a) Dark-room which is the ideal environment. (b) Lights for Fig. 2.1(b) with low inter-reflection. (c) Lights for Fig. 2.1(c) with high inter-reflection. (d) Challenging outdoor scene for Fig. 2.1(d).

Example	Image Size	# of images	Standard Running time	Coarse-to-fine Running time
Syn. Case	101×101	9	49.7s	0.16s
KitCat	357×442	14	1499.1s	7.02s
HumanDoll	299×714	10	253.1s	4.84s
Kitten	415×666	12	538.2s	8.33s
WineBottle	548×870	8	2048.6s	7.23s
Vase	268×576	12	285.9s	5.45s

Table 2.1: Running time for the examples shown in this thesis. It is reported on a desktop computer with a Core2Duo CPU 2.79GHz and 2GB memory.

as $(R = (u_x+1)/2, G = (u_y+1)/2, B = u_z)$ from surface normal $(u_x, u_y, u_z)^T$. The final surfaces are constructed using the method in [2].

Toy Example We synthesized a set of images for a sanity check. Fig. 2.10 shows an example, where the input 9 images are rendered with environment lighting. With the ground truth surface normal in (c), which is a hemisphere, we compare the results produced using traditional PS methods [41] without considering environment lighting and our new approach. They are respectively shown in images (d)-(f) and images (g)-(i). It is noticeable that ignoring the environment lighting flattens the resulting surfaces. The average angular errors (AAE) are 28.85 and 0.39 degrees respectively, differing hugely.

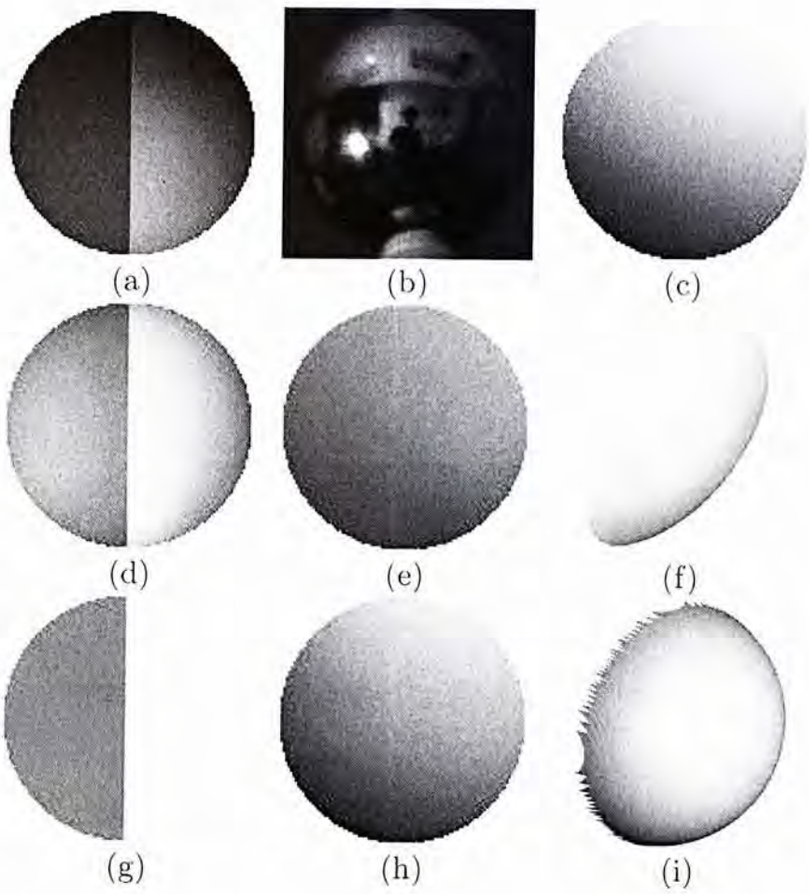
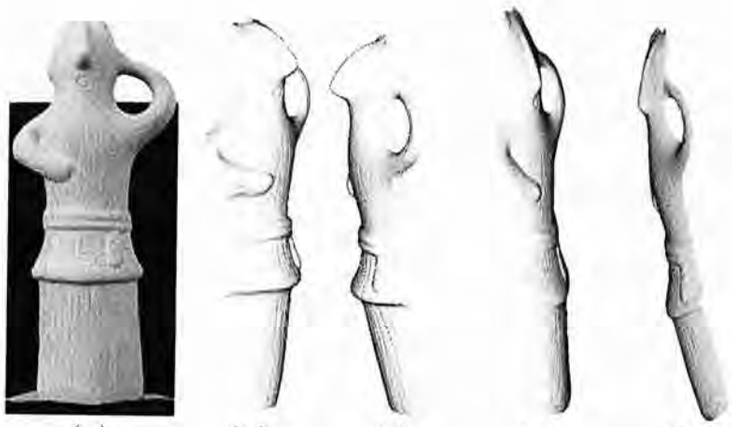


Figure 2.10: *Synthetic Case* – (a) One out of 9 input images. (b) The corresponding mirror sphere. (c) The ground truth normal map (color coded). (d)-(f) are the albedo and normal maps, and the reconstructed surface by only finding the major light sources. (g)-(i) show our results, very close to the ground truth.

Different Environment Evaluation In this example, we put the *KitCat* in different challenging environments shown in Fig. 2.1(a)-(c) for surface reconstruction using PS. The object surface in Fig. 2.12(a) contains glossy reflection. In all of the environments, we test either using a handheld lamp to illuminate the object or switching on and off different ceiling fluorescent tubes to produce photometric parallax. Our quantitative evaluation shown in Table 2.2 by comparing to the ground truth result obtained in a dark-room indicates that our method can generally produce decent normal maps and surfaces in these situations. The average angular errors are unexceptionally



(a) (b) (c) (d) (e)

Figure 2.11: Outdoor example. (a) An input image. (b)-(c) Two views of the reconstructed surface using our method. (d)-(e) Two views of the surface result by the traditional method [41].

Environment	Surface Normal AAE
(a) Dark-room	3.234815
(b) Low inter-reflection	5.621666
(c) High inter-reflection	8.134300

Table 2.2: Average angular errors (AAE) of the estimated normal maps, corresponding to the results shown in Fig. 2.12(d)-(f).

small even with cast shadow influence in the input images.

Outdoor Scene Fig. 2.11 shows an outdoor example *HumanDoll* with the environment shown in Fig. 2.1(d). We simply put the experimental equipment on a trolley and move it to change incoming lights without any human-controlled lighting. We captured two sets of images with and without the presence of the sun. The latter was achieved under eaves. All these images are input to our system. Our result is with high quality compared to the one that finds the major light source without considering the environment.

More Examples Fig. 2.13 shows the results of *Kitten*, *Vase*, and *WineBottle*. For the *Kitten* example environment, the sun shines into the room though windows. This discontinuous extention of "ambient" is not allowed in other

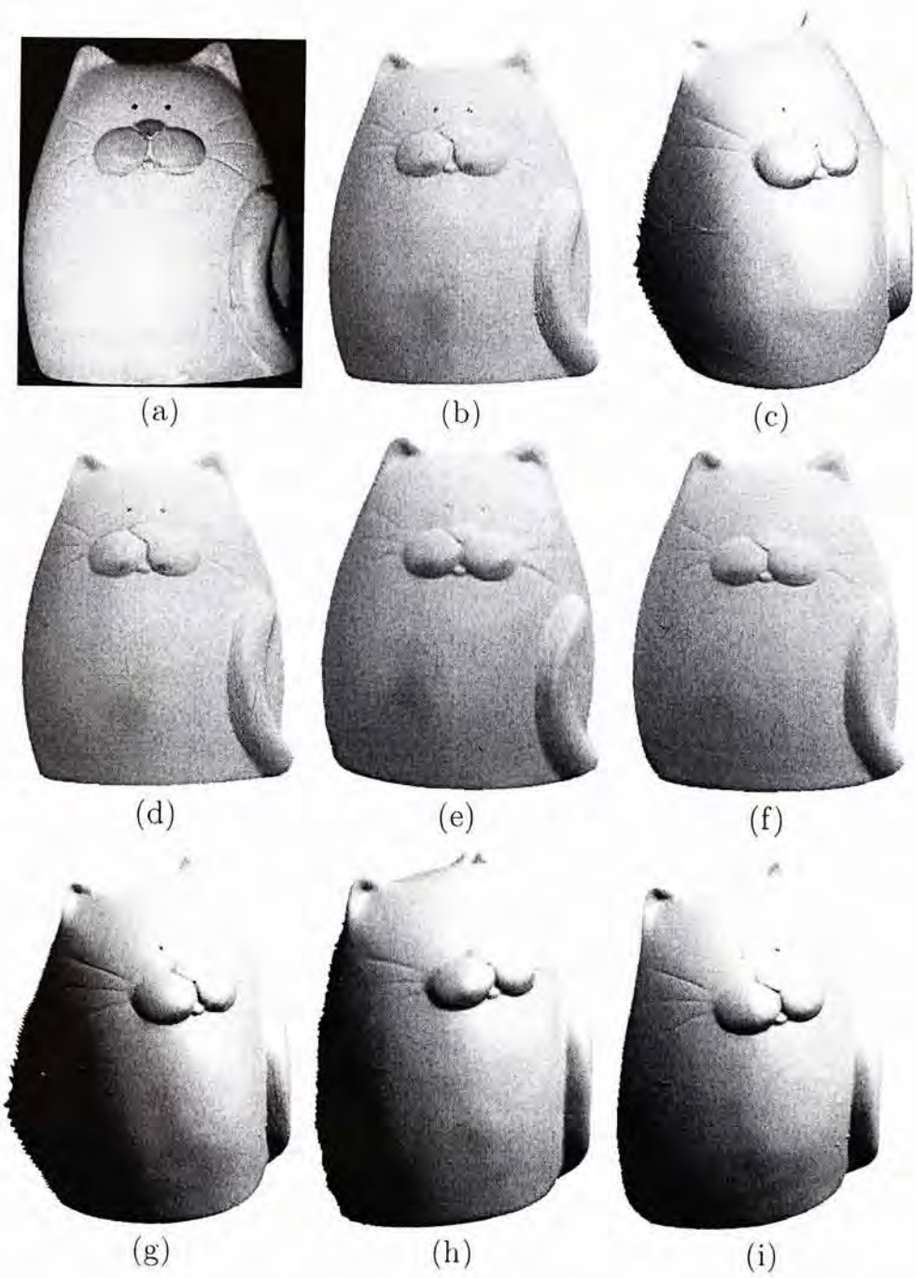


Figure 2.12: *KitCat* – (a) An input image. (b) Ground truth normal map generated in a dark-room. (c) The reconstructed surface from (b). (d)-(f) Color coded normal maps generated by our method with the environments respectively shown in Fig. 2.1(a)-(c). (g)-(i) Corresponding surfaces.

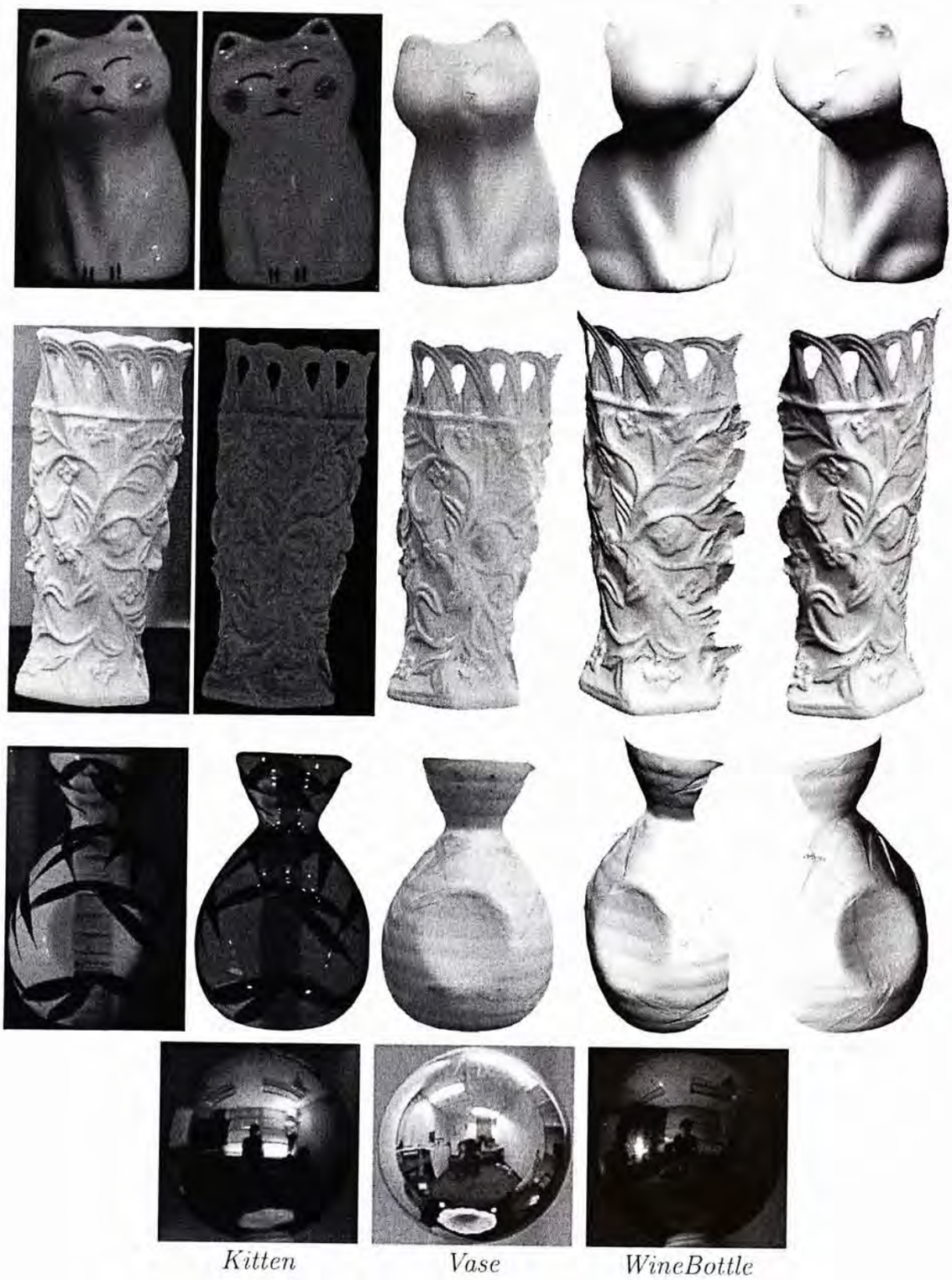


Figure 2.13: More examples. From top to bottom are *Kitten*, *Vase*, *WineBottle* and the corresponding mirror spheres. From left to right (for top 3 rows) – an input image, surface albedo, color coded normal map, surface results in two views.

PS methods. The *Vase* is captured in a highly reflective room. Decent surface normal and albedo are produced for all examples. Important structure and subtle details on the surfaces are also properly maintained.

□ **End of chapter.**

Chapter 3

Generalized Stereo Matching

3.1 Problem Description

With the precipitate prevalence of 3D display and 3D capturing devices, a tremendous number of binocular videos come into existence. If depth can be accurately computed in them with necessary temporal consistency, challenging video editing to alter color, structure, and geometry, as well as the high-level scene understanding and recognition tasks can be achieved more easily. Nonetheless, with only two views, it is very difficult to compute reliable geometry information in long sequences.

In a binocular video that contains moving or deforming objects, structure-from-motion (SFM) and multi-view stereo matching cannot be employed because of the violation of the multi-view geometry, handicapping correspondence establishment in multiple frames. 2D optical flow together with the depth variation are jointly considered in [29, 43, 38, 21, 37], typically referred to as the scene flow field, for 3-dimensional displacement estimation. These methods either compute the motion and depth independently or resort to a four-image configuration. They do not tackle the temporal-consistency

preservation problem in multiple frames and are not suitable to handle long sequences.

We propose a new method for reliable depth and motion estimation from multi-frame binocular videos, and aim to faithfully maintain dense temporal-consistency in the results, especially along discontinuous boundaries. This problem was generally regarded as very difficult without the global multi-view geometry constraint, and was seldom addressed in prior work, counting in dynamic objects with varying boundaries, image noise, frequent occlusion, and other visual artifacts. Image-pair-based estimation only yields locally optimal results. In a long sequence with many frames, consecutively applying this scheme easily accumulates errors, making estimates quickly deviate from the correct values. Worse, there is no effective method to identify and correct the estimation errors in multiple frames.

Our method tackles all aforementioned difficulties without making any static-object assumption. We make several major contributions to construct a robust system that has the ability to spot and rectify estimation errors in multiple frames. 1) First, we propose the *motion trajectory* that links reliable motion corresponding points among frames. It is robust against occasional noise and lighting artifacts. When temporally-consistent occlusion arises, the trajectories can be automatically broken to avoid outliers. 2) We build structure profiles by simultaneously considering multi-frame edges. Through a voting-like average step, only edges reliable in multiple frames are enhanced. 3) To correct disparity and motion boundary values, we also introduce *trajectory*-based estimate confidence, constraining strictly the disparity and motion data. 4) Last but not the least, we propose anisotropic smoothing functions to non-uniformly regularize pixels, where possible edges are required to be smooth only along the isophote direction, preventing un-

constrained boundary generation.

3.2 Related Work

Simultaneous depth and motion estimation from stereo images was studied in [44]. Following it, methods of [44, 29, 39] computed motion and depth sequentially and independently, assuming that the depth in previous frames is known. To improve the result quality, depth and motion are jointly estimated, using two stereo pairs [43, 25, 21, 37]. These approaches compute two consecutive depth maps each time. Note that it is difficult to apply them sequentially to long sequences without an effective error rectification scheme.

Efforts have also been put to motion/depth discontinuity handling. Zhang and Kambhamettu [43] used segmentation and applied piecewise regularization. Edge-preserved regularizer [25] and complementary regularizer [37] were used to preserve motion and depth boundaries. The color edges or segments that are used as guidance are generally hard to be consistent over time, making producing high-quality depth-boundary difficult.

Correspondence of points in a dynamic scene is typically established using optical flow [21]. Besides the rapid development of two-frame methods [8, 10, 45, 42], several approaches can be applied to enforce temporal smoothness. In [8, 10], temporal smoothness is yielded by assuming that the flow vectors from consecutive frames at the same image location are similar. It applies to smoothly-varying motion. Álvarez *et al.* [4] enforced symmetry between the forward and backward flow vectors to reject outliers. This method does not tackle as well the temporal consistency problem in video sequences. Using multiple frames, Irani [22] projected flow vectors onto a subspace and assumed that the resulted matrix has a low rank for noise removal. Rigidly moving objects are considered in this approach. To construct long-range

motion trajectories, particle samples were generated and linked in a royalistic way [31]. Contrary to all these approaches, in this thesis, we propose a general binocular framework addressing the temporal consistency problem in depth and motion estimation in long sequences. No restriction on object motion is imposed.

3.3 Our Approach

Our approach contains a few important steps. We consecutively initialize motion and depth in two views respectively, construct reliable motion trajectories for all pixels, form voting-like edge profile, develop depth/motion volume priors, and finally define well-regularized object functions over multiple frames to compute temporally consistent depth and motion.

3.3.1 Notations and Problem Introduction

With the input rectified binocular video, our method computes the two-view stereo information for each corresponding frame pair across the two sequences, together with dense motion maps in each sequence. Our framework is general and can be readily extended to unrectified stereo videos linked with a fundamental matrix [37].

We denote corresponding frames in the stereo sequences as f_l^t and f_r^t, indexed by time t where $t = \{0, 1, ...n-1\}$. For each pixel x in frame f_l^t, we find the corresponding pixels in other neighboring frames either temporally using motion estimation or spatially with stereo matching, as shown in Fig. 3.1. The correspondence x_r^t in f_r^t can be written as

$$x_r^t = x + \mathrm{d}_l^t,$$

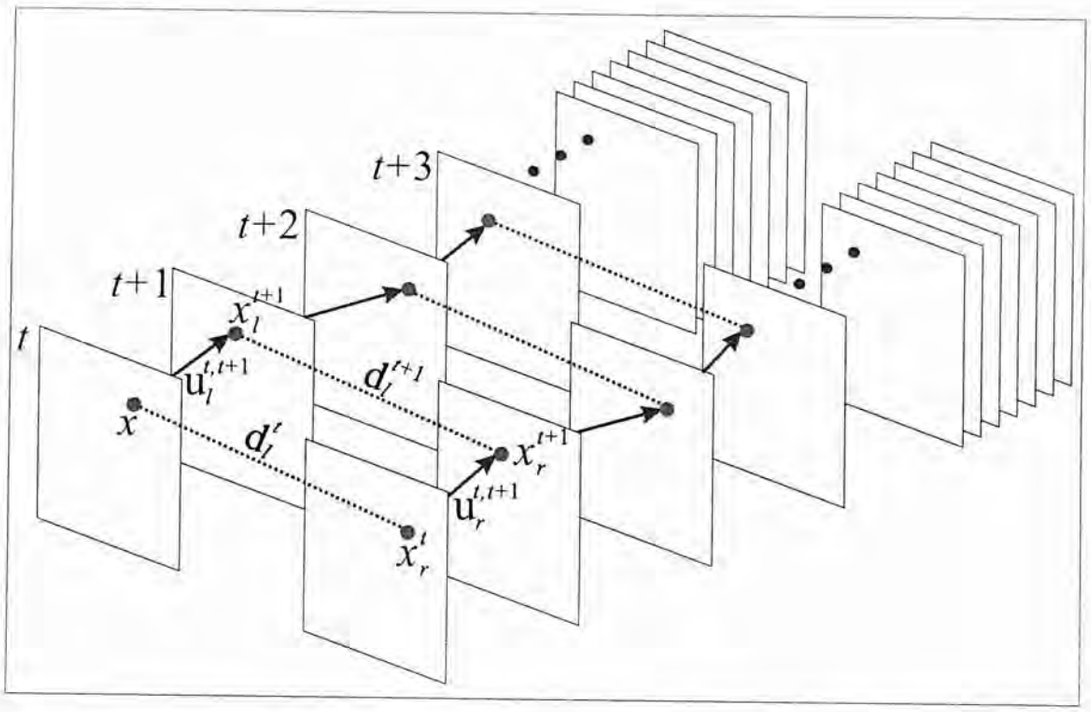


Figure 3.1: Correspondences of x in different frames.

where d_l^t is the disparity of x. Meanwhile, by optical flow estimation in the left sequence, we can find its motion correspondence

$$x_l^{t+1} = x + \mathrm{u}_l^{t,t+1}$$

in the temporally neighboring frame f_l^{t+1}, where the optical flow 2D vector $\mathrm{u}_l^{t,t+1}$ is written as $\mathrm{u}_l^{t,t+1} = (u_l^{t,t+1}, v_l^{t,t+1})^T$, as shown in Fig. 3.1. The correspondence in f_r^{t+1} is $x_r^{t+1} = x + \mathrm{u}_l^{t,t+1} + \mathrm{d}_l^{t+1}$, involving both the motion and stereo matching results.

With these correspondences, we can establish important constraints for motion and depth estimation. The following two terms are used in our method

$$\begin{aligned} E_{D1} &= \Gamma(|f_r^t(x + \mathrm{d}_l^t) - f_l^t(x)|^2), \\ E_{D2} &= \Gamma(|f_l^{t+1}(x + \mathrm{u}_l^{t,t+1}) - f_l^t(x)|^2), \end{aligned} \tag{3.1}$$

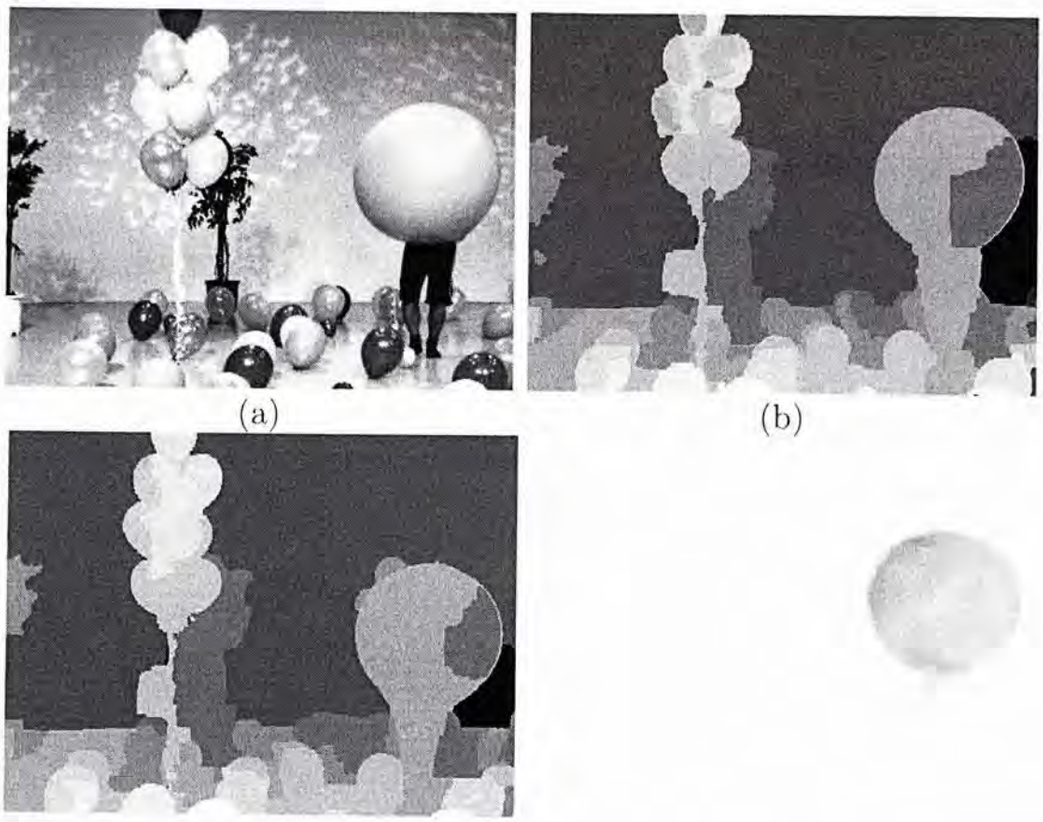
where $\Gamma(\cdot)$ is the robust Charbonnier function, written as $\Gamma(y^2) = \sqrt{y^2 + \epsilon^2}$ to reject matching outliers. In Eq. (3.1), each pixel takes a 5-channel input,

i.e., $f = (f_R, f_G, f_B, f_{\partial h}, f_{\partial v})$, robust against illumination variation [31]. $f_{\partial h}$ and $f_{\partial v}$ are intensity gradients projected to the horizontal and vertical axes.

Simultaneously minimizing these motion estimation and stereo matching terms along with regularization functions in *all* frames is computationally intractable. It is also not reliable even with only 10 frames because all correspondences are *locally* established, each between only two images, which is vulnerable to noise and illumination variation. Unlike multi-view stereo, there is no global constraint to reduce errors over the sequences. Other schemes to sequentially computing depth maps are also not proper. When one depth estimate is problematic, all following computation steps will be affected, inevitably accumulating errors and eventually failing the process. Therefore, effective methods are required to rectify estimation errors or skip outliers in the middle of computation. We propose new trajectory-based structure priors to accomplish this goal.

3.3.2 Depth and Motion Initialization

Our method first initializes the variables. Although Eq. (3.1) contains similar forms for finding depth and motion flow correspondences, they are inherently different. Depth estimation by two-view stereo matching is generally achieved with discrete optimization due to possibly large disparities panning many pixels. Optical flow estimation, on the contrary, is developed for consecutive frame motion estimation where small displacements are typically resulted. In our initialization step, discrete and continuous optimizations are employed respectively to compute disparity and flow. These values will be updated later with the effective temporal constraints.



(a) (b)

(c) (d)

Figure 3.2: Initial depth and flow estimates. (a) One frame in the left-view sequence. (b)-(c) The depth maps for two consecutive frames. (d) One optical flow field.

Disparity Initialization We compute the disparity d_l^t and d_r^t for each $t \in \{0, 1, ...n-1\}$ by optimizing

$$E(d^t) = E_{D1}(\mathrm{d}^t) + \beta_1 E_{S1}(\nabla \mathrm{d}^t), \tag{3.2}$$

where $E_{S1}(\nabla \mathrm{d}^t) = \min((\nabla \mathrm{d}^t)^2, \rho)$ is the truncated quadratic function for preserving discontinuities, with $\rho = 10$. It is a regularization term. β_1 is a weight set to 20. The functions are solved by graph-cuts [23], which deal with large disparity and occlusion. Two depth maps for consecutive frames are shown in Fig. 3.2(b)-(c). Textureless regions and region boundary pixels could have inconsistent estimates, which will be refined in the following steps.

Initial Optical Flow Estimation We initialize 2D motion field in a variational framework, which is capable of capturing sub-pixel motion between two consecutive frames. Taking the left view for example, we compute both the forward and backward flow vectors, denoted as $\mathrm{u}_l^{t,t+1}$ and $\mathrm{u}_l^{t+1,t}$, for outlier rejection. For robust estimation in the main steps, which will be detailed in Sec. 3.3.3, we also compute bi-directional optical flow between frames f_l^t and f_l^{t+2} (denoted as $\mathrm{u}_l^{t,t+2}$ and $\mathrm{u}_l^{t+2,t}$), and between frames f_l^t and f_l^{t+3} (denoted as $\mathrm{u}_l^{t,t+3}$ and $\mathrm{u}_l^{t+3,t}$). Fig. 3.3 illustrates these vectors. As all motion vectors are computed similarly, in what follows, we only describe estimation of $\mathrm{u}_l^{t,t+1}$. The same initialization process repeats in the right view.

Forward optical flow $\mathrm{u}_l^{t,t+1}$ is computed by minimizing

$$E(\mathrm{u}_l^{t,t+1}) = E_{D2}(\mathrm{u}_l^{t,t+1}) + \beta_2 E_{S2}(\nabla \mathrm{u}_l^{t,t+1}), \tag{3.3}$$

where $E_{S2}(\nabla \mathrm{u}_l^{t,t+1})$ is the total variation regularizer, expressed as $\sqrt{|\nabla \mathrm{u}_l^{t,t+1}|^2 + \epsilon^2}$ to preserve edges. It is commonly used in the variational setting. β_2 is a weight, uniformly set to 12. Eq. (3.3) is optimized by the efficient method of [8].

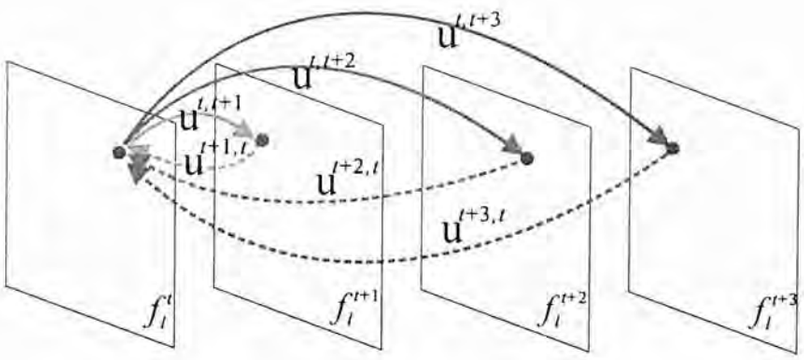


Figure 3.3: Flow vector illustration.

3.3.3 Volume-based Structure Prior

With the initial depth maps and motion fields, we construct motion trajectories for all pixels, which link among frames corresponding pixels. They are then used to define edge profiles, essential in our system to form new priors to update estimates.

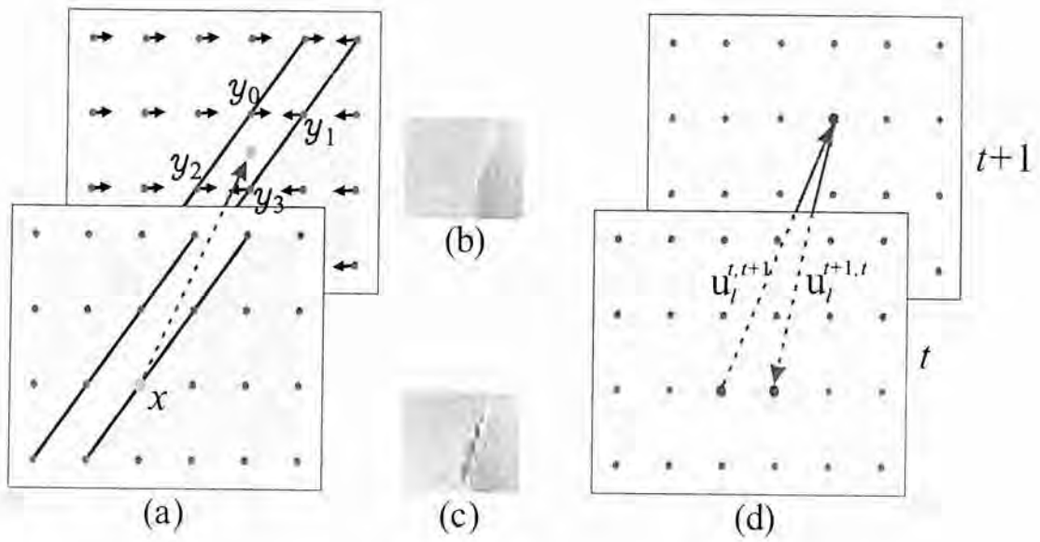


Figure 3.4: Illustration of (a) flow interpolation problem, (b) accumulated motion and backward warping for 7 continues frames using our bilateral interpolation, (c) accumulated motion and backward warping using bilinear interpolation, (d) bidirectional consistency check.

Robust Motion Trajectory Forming motion trajectories among corresponding pixels in each sequence needs to link sub-pixel motion vectors. For ease of description, we do not specify the left or right sequence and omit the index. Specifically, $x + \mathrm{u}^{t,t+1}(x)$ in f^{t+1} that is mapped from x in f^t based on the motion vector $\mathrm{u}^{t,t+1}(x)$ is possibly a fractional value, locating in between four pixels, as shown in Fig. 3.4(a). So its actual motion vector has to be estimated by interpolation.

Simple distance-based interpolation, e.g., bilinear or bicubic method, is

not the best strategy. Fig. 3.4(a) shows one example that the orange and green regions undergo motion in different directions. A point projected in between pixels $\{y_0, y_1, y_2, y_3\}$, after bilinear interpolation, is with near zero motion magnitude, which is obviously impropriate. To improve the quality, color difference is also considered in our system to guide interpolation bilaterally together with the spatial distance, which yields the following operator:

$$\mathrm{u}^{t+1,t'}(x+\mathrm{u}^{t}(x)) = \frac{1}{|w|}\sum_{i=0}^{3}\mathrm{u}^{t+1,t'}(y_i)\cdot e^{-|x-y_i|^2/\gamma_1-|f^{t'}(x)-f^{t+1}(y_i)|^2/\gamma_2}, \tag{3.4}$$

where t' can be t, $t+2$, or other frame indexes depending on the motion definition. $|w|$ is for normalization. The term $|f^t(x)-f^{t+1}(y_i)|^2/\gamma_2$ considers the color similarity of points in different frames. The performance of the bilateral interpolation and bilinear interpolation are compared in Fig. 3.4(b)-(c).

Using this interpolation method, we link corresponding points among frames, which forms *motion trajectories*. Due to some initial motion estimation errors especially in occlusion, object boundary, and textureless regions, we identify incredible estimates and exclude them in trajectory construction based on the computed bidirectional flow vectors. Specifically, we project $x+\mathrm{u}^{t,t+1}(x)$ in f^{t+1}, which is mapped from x in f^t based on the motion vector $\mathrm{u}^{t,t+1}(x)$, back to f^t. We sum the two vectors with opposite directions and check if amplitude

$$|\mathrm{u}^{t+1,t}(x+\mathrm{u}^{t,t+1}(x))+\mathrm{u}^{t,t+1}(x)| \geq 1.$$

Satisfying the inequality means the motion vectors that are supposedly opposite contain large errors. We thus discard $\mathrm{u}^{t,t+1}(x)$. One example is shown in Fig. 3.4(d), where the flow vector $\mathrm{u}^{t,t+1}(x)$ is regarded as wrong.

Removing a problematic flow vector breaks a link into two. Doing it too often in a sequence will result in many short motion trajectories, which is undesirable. We notice that many incorrect vectors are actually caused by sudden noise and lighting outliers. They in general do not appear consecutively in several frames for the same pixel.

To make our method robust against this type of noise inference, we make use of the higher-range bidirectional flow vectors, i.e., $\mathbf{u}_l^{t,t+2}(x)$ and $\mathbf{u}_l^{t+2,t}(x)$, produced in (3.4). If they are found reliable after going through the same bidirectional consistency check, we reconnect the trajectory from frame t to $t+2$. Otherwise, we continue to test the pair of $\mathbf{u}_l^{t,t+3}(x)$ and $\mathbf{u}_l^{t+3,t}(x)$. Only if all these three checks fail, we break the trajectory.

For the example shown later in Fig. 3.9, we used a 50-frame sequence and found that the average lengths of the motion trajectories are 7 and 29 respectively when using the one-pair-vector and three-pair-vector consistency check. The random noise influence is greatly reduced in the latter case. It is noteworthy that consistent occlusion can break links. This result is desirable because unreliable correspondences will be removed from trajectories.

Trajectory-based Structure Profile Unlike multi-view stereo that has the global geometry information to infer depth, in motion estimation, good constraint that can be applied across multiple frames is hard to find. We present a temporal voting-like scheme based on a key observation – that is, salient object boundaries are more distinctive than other pixels and typically vary coherently in multiple frames. Our new method can find those boundary pixels and establish strong-edge confidence for multi-view motion refinement.

We in the first place calculate edge magnitude maps from bilateral-filtering applied images to remove a small degree of noise. Each gradient magnitude is written as $\sqrt{f_{\partial h}^2 + f_{\partial v}^2}$. Our *edge occurrence* map C_f^t is obtained by set-

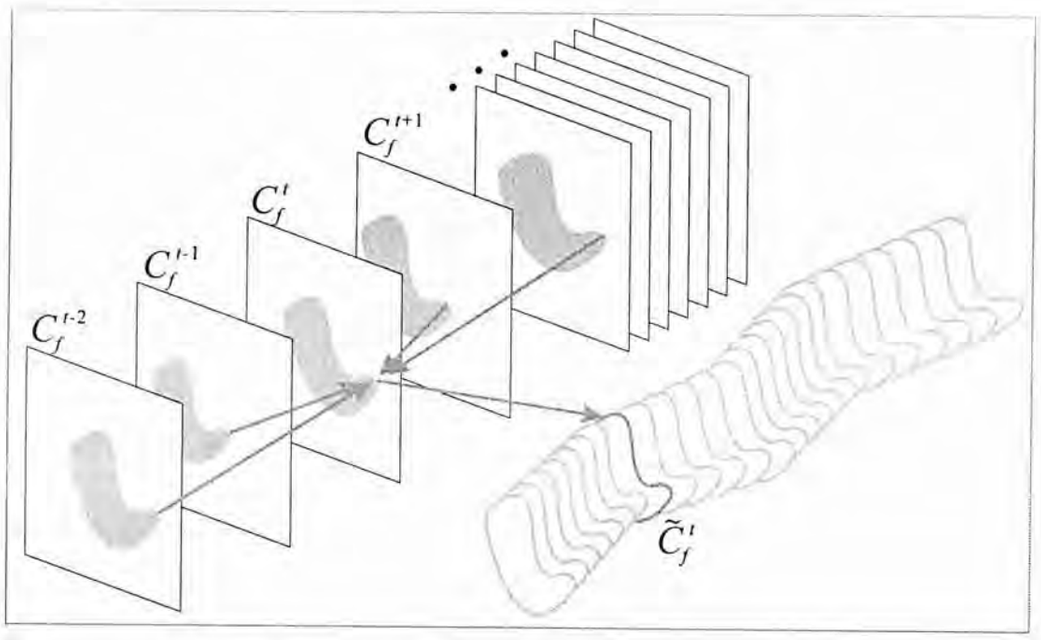


Figure 3.5: Trajectory-based structure profile construction.

ting pixels with their magnitudes smaller than a threshold (generally set to 0.01) to zeros. These maps in a sequence, together with the computed dense motion trajectories, are used to establish a structure profile map for each frame.

Based on the corresponding points in motion trajectories, for each pixel x in frame t, assuming it is in trajectory j, we project all *edge occurrence* values of the points in trajectory j to it, as illustrated in Fig. 3.5. This process can find consistent edge occurrence values where occasional errors after averaging can be quickly suppressed.

The corresponding point of x in frame $t+i$ is $x+\mathrm{u}^{t+i,t}(x)$ after chain projection, where $\mathrm{u}^{t+i,t}(x)$ is the motion vector. The average of the occurrence value in the trajectory is expressed as

$$\tilde{C}_f^t(x) = \frac{1}{n_j} \sum_i C_f^{t+i}(x + \mathrm{u}^{t+i,t}(x)), \tag{3.5}$$

where n_j is the length of the trajectory j. The average process is a robust voting-like scheme that gathers confidence to reduce fitful errors that cannot

be continuous and consistent in several frames. So the sum of errors is generally far smaller than the correct estimates.

The resulted view-dependent average edge occurrence maps are $\{\tilde{C}_f^0, \tilde{C}_f^1, ...\tilde{C}_f^t, ...\tilde{C}_f^{n-1}\}$ which are used to define edge priors. Fig. 3.5 shows one illustration. The 4^{th} row of Fig. 3.9 shows the averaged edge occurrence maps, where inconsistent edges are notably weakened. This map construction process can enhance edges even when they are influenced by noise and illumination change in the current frame.

Trajectory-based Depth/Motion Profile $\tilde{C}$s contain reliable structures for further depth refinement. To improve the quality, we also apply the bilateral flow filtering method [31] to them, which performs smoothing with respect to weights defined with spatial proximity, color similarity, motion similarity, and reliability labeling. Reliability labeling is achieved in our system using bidirectional checking. The filtering method is applied to regions within 10 pixels around edges in map $\tilde{C'}_f$. Edge pixels are decided if their values are larger than 0.05 in a $[0, 1]$ scale.

In this step, we also compute profiles for further depth and motion magnitude regularization because $\tilde{C}_f$ does not contain actual magnitude information. It is done by weighted average of corresponding depth values or motion magnitude in each trajectory, using a temporally weighted Gaussian window. For depth, it is expressed as

$$\tilde{\mathrm{d}}^t(x) = \frac{1}{|w_d|} \sum_i \mathrm{d}^{t+i}(x + \mathrm{u}^{t+i,t}(x)) e^{-i^2/\sigma_t}, \tag{3.6}$$

where σ_t is the standard deviation, set to 20. Frame i is inside the corresponding trajectory. $|w_d|$ is for normalization. The resulted depth volume is denoted as $\{\tilde{\mathrm{d}}_l^0, \tilde{\mathrm{d}}_l^1, ...\tilde{\mathrm{d}}_l^t, ...\tilde{\mathrm{d}}_l^{n-1}\}$.

The flow volume can be similarly computed with a smaller σ_t, set to 5.

The motion volume is denoted as $\{\tilde{\mathrm{u}}_l^0, \tilde{\mathrm{u}}_l^1, ...\tilde{\mathrm{u}}_l^t, ...\tilde{\mathrm{u}}_l^{\prime\prime-1}\}$. Frames in the depth and motion volumes are consistent compared to the initial d and u.

3.3.4 Objective Function with Volume-based Priors

The edge, motion, and depth profiles can greatly help depth and motion refinement due to its fidelity and consistency to define strong edges and magnitudes among frames, describing high quality depth-motion boundary shapes. Our depth estimation is based on minimizing

$$E'(\mathrm{d}_l^t) = E_{D1}(\mathrm{d}_l^t) + \beta_3 E_{T1}(\mathrm{d}_l^t; \tilde{d}_l^t) + \beta_4 E_{T2}(\nabla \mathrm{d}_l^t; \tilde{C}_f^t) \tag{3.7}$$

where E_{T1} is the depth value prior and E_{T2} models edge occurrence confidence. β_3 and β_4 are weights set to 20 and 12. E_{T1} is defined as

$$E_{T1}(\mathrm{d}_l; \tilde{\mathrm{d}}_l) = (\mathrm{d}_l - \tilde{\mathrm{d}}_l)^2, \tag{3.8}$$

requiring that the computed depth values do not vary too much from $\tilde{\mathrm{d}}_l$.

Unlike the four-image configuration for scene flow estimation [21, 37], Eq. (3.7) involves depth estimates for two views and the edge prior actually encodes the key structural information across multiple frames temporally. We enforce smoothness for regions with small edge-occurrence values in $\tilde{C}_f^t$ and allow depth discontinuity to take place when $\tilde{C}_f^t(x)$ is large. On account of possible errors in averaging the edge-occurrence maps, which make some edges in $\tilde{C}_f^t$ slightly wider than what they should be (at most 2-4 pixels wider in our experiments), it is impropriate to naively enforce no or small smoothness for the edge pixels in $\tilde{C}_f^t$. Without necessary regularization, edge discontinuity result will be unpredictable.

We turn to an *anisotropic* smoothness method to provide critical constraint for edge-preserving regularization. We first decompose depth gradient

∇d into $\{\nabla \mathrm{d}^{\parallel}, \nabla \mathrm{d}^{\perp}\}$ according to the image gradient direction, where

$$\nabla d^{\parallel} = \langle \nabla \mathrm{d}, \frac{\nabla f}{\|\nabla f\|} \rangle, \nabla \mathrm{d}^{\perp} = \nabla \mathrm{d} - \nabla \mathrm{d}^{\parallel}.$$

∇f is the corresponding frame intensity gradient. We propose the following function for smoothness regularization:

$$E_{T2}(\nabla \mathrm{d}; \tilde{C}_f) = \Gamma((\nabla d^{\perp})^2 + (1 - \tilde{C}_f)(\nabla d^{\parallel})^2), \tag{3.9}$$

where $\Gamma(\cdot)$ is the robust Charbonnier function, defined in Eq. (3.1). In Eq. (3.9), for all pixels, smoothness is enforced along the isophote direction. We also allow discontinuity along gradient only for reliable strong-edge pixels, which corresponds to large values in $\tilde{C}_f$. Hence, Eq. (3.9) provides necessary constraints in different types for the pixels. With a few algebraic operations, Eq. 3.9 can be written in a form of diffusion tensor

$$E_{T2}(\nabla \mathrm{d}; \tilde{C}_f) = \Gamma(\nabla \mathrm{d}^T D(\nabla f) \nabla d), \tag{3.10}$$

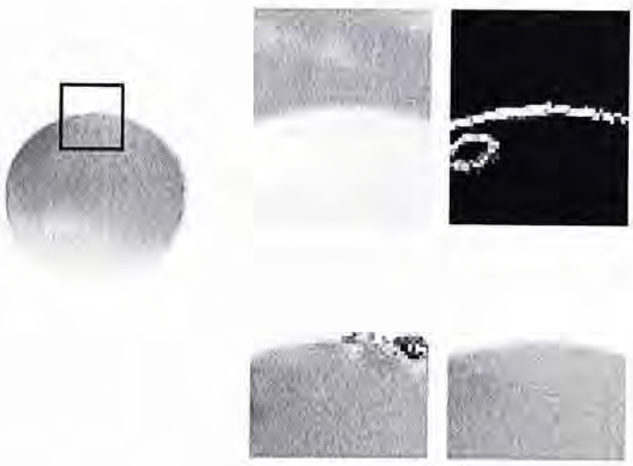
where $D(\nabla f)$ is the diffusion tensor defined as

$$D(\nabla f) = \frac{1}{\|\nabla f\|} \left((\nabla f^{\perp})(\nabla f^{\perp})^T + (1 - \tilde{C}_f)(\nabla f^{\parallel})(\nabla f^{\parallel})^T \right).$$

To minimize Eq. (3.7), we use the variational method. Based on the fact that depth d_l generally does not deviate too much from the prior $\tilde{d}_l$, we perform linearization at $\tilde{d}_l$ and compute the increments Δd_l. Eq. (3.8) is accordingly changed to

$$E_{T1} = \Delta d_l^2,$$

which is the Tikhonov regularization on Δd_l, indicating that our temporal prior globally constrains the depth estimates by allowing smooth variation. The final estimation result is $d_l = \tilde{d}_l + \Delta d_l$.



(a) (b)

Figure 3.6: Regularization effectiveness. (a) Our optical flow estimate. (b) From top to bottom and from left to right: patch of input image f^t, $\tilde{C}_f^t$, optical flow estimates using isotropic and anisotropic regularization terms.

The objective function to update motion is similarly defined, expressed as

$$E'(\mathrm{u}_l^{t,t+1}) = E_{D2}(\mathrm{u}_l^{t,t+1}) + \beta_3 E_{T1}(\mathrm{u}_l^t; \tilde{\mathrm{u}}_l^t) + \beta_4 E_{T2}(\nabla \mathrm{u}_l^t; \tilde{C}_f^t). \quad (3.11)$$

Eqs. (3.7) and (3.11) are in the same form except for the data term, which uses different image pairs for matching. The following numerical solution is proposed to minimize Eq. (3.11). Eq. (3.7) can be optimized similarly, in a simpler manner in only 1D.

One example of optical flow using the temporal prior is shown in Fig. 3.6(a), with the comparison in (b). The bottom left subfigure of (b) shows the flow estimates obtained by enforcing smoothness uniformly with $(1 - \tilde{C}_f^t)$. When $\tilde{C}_f^t$ is large near the boundaries, the flow estimation is ill-posed, resulting in a problematic field. Our result with anisotropic regularization is also presented.

3.3.5 Numerical Solution

The variational function (3.11) is solved with the Euler-Lagrange equations. Linearization is performed on the data term using the Taylor expansion. It solves for the increment $\Delta \mathrm{u} = \mathrm{u} - \tilde{\mathrm{u}}$.

For ease of description, we denote $f_h = \partial_h f^{t+1}(x+\tilde{u}(x))$, $f_v = \partial_v f^{t+1}(x+\tilde{u}(x))$, $f_z = f^{t+1}(x+\tilde{u}(x)) - f^t(x)$, and $b = f_h \cdot \Delta u + f_v \cdot \Delta v + f_z$. Γ'_d is used to represent $\Gamma'(b^2)$, the derivative of robust function, and $\Gamma'_s = \Gamma'(\nabla u^T D(\nabla f)\nabla u + \nabla v^T D(\nabla f)\nabla v)$. The Euler-Lagrange equations for Eq. (3.11) are given by

$$\Gamma'_d \cdot b f_h + 2\beta_3 \Delta u - \beta_4 \mathrm{div}(\Gamma'_s \cdot D(\nabla f)\nabla u) = 0,$$
$$\Gamma'_d \cdot b f_v + 2\beta_3 \Delta v - \beta_4 \mathrm{div}(\Gamma'_s \cdot D(\nabla f)\nabla v) = 0,$$

where div is the divergence operator. A fixed-point loop similar to that in [8, 9] is applied, which removes the nonlinearity of Γ'. Then the equation system is solved by the coupled point Gauss-Seidel relaxation [9]. With the applied anisotropic diffusion tensor, the smoothness term involves several neighboring points. We use the indices in Figure 3.7 to represent the 2D coordinates, e.g., $u_1 = u(i+1, j+1)$. q is used to index the current point $x = (i, j)$.

By defining $r_h = \frac{f_{\partial v}^2 + (1-\bar{C}) f_{\partial h}^2}{|\nabla f|^2 + \epsilon}$, $r_v = \frac{f_{\partial h}^2 + (1-\bar{C}) f_{\partial v}^2}{|\nabla f|^2 + \epsilon}$, $r_{h1} = r_h(i+1, j+1)$, and $r_c = \frac{-\bar{C} f_{\partial h} f_{\partial v}}{|\nabla f|^2 + \epsilon}$, we represent the anisotropic factors in simpler forms. We discretize a grid with the size $h_h \times h_v$ to apply Gauss-Seidel relaxation, which is written as

$$\begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} \\ m_{12} & m_{22} \end{pmatrix}^{-1} \begin{pmatrix} u_r \\ v_r \end{pmatrix}, \tag{3.12}$$

7 (i-1,j+1)	0 (i,j+1)	1 (i+1,j+1)
6 (i-1,j)	q (i,j)	2 (i+1,j)
5 (i-1,j-1)	4 (i,j-1)	3 (i+1,j-1)

Figure 3.7: Indices of the 2D coordinates.

where

$$m11 = \Gamma'_d \cdot (f_h)^2 + 2\beta_3 + \sum_{\diamond \in \{h,v\}} \sum_{p \in \mathcal{N}_\diamond(q)} \beta_4 \frac{\Gamma'_{sq} r_{\diamond q} + \Gamma'_{sp} r_{\diamond p}}{2h_\diamond^2},$$

$$m12 = \Gamma'_d \cdot f_h f_v.$$

$$m22 = \Gamma'_d \cdot (f_v)^2 + 2\beta_3 + \sum_{\diamond \in \{h,v\}} \sum_{p \in \mathcal{N}_\diamond(q)} \beta_4 \frac{\Gamma'_{sq} r_{\diamond q} + \Gamma'_{sp} r_{\diamond p}}{2h_\diamond^2},$$

$\mathcal{N}$ is the set of neighboring pixels, $\mathcal{N}_h(q) = \{2, 6\}$, and $\mathcal{N}_v(q) = \{0, 4\}$. Through a few algebraic operations, the u_r and v_r in (3.12) can be derived as

$$u_r = -\Gamma'_d \cdot f_h f_z + g(u),$$

$$v_r = -\Gamma'_d \cdot f_v f_z + g(v),$$

where

$$g(a) = \sum_{\diamond \in \{h,v\}} \sum_{p \in \mathcal{N}_\diamond(q)} \beta_4 \frac{\Gamma'_{sq} r_{\diamond q} + \Gamma'_{sp} r_{\diamond p}}{2h_\diamond^2} a_q + \sum_{p \in \{0,2,4,6\}} \beta_4 \frac{\Gamma'_{sr} r_{cp} + \Gamma'_{sq} r_{cq}}{2h_h h_v} \frac{(a_{\overline{p-2}} + a_{\overline{p-1}} - a_{\overline{p+1}} - a_{\overline{p+2}})}{4}.$$

$\overline{p} = p \mod 8$. The Gauss-Seidel relaxation practically runs in three iterations, which are enough to yield a good Δu result. Depth increment Δd can be computed in a similar procedure in 1D.

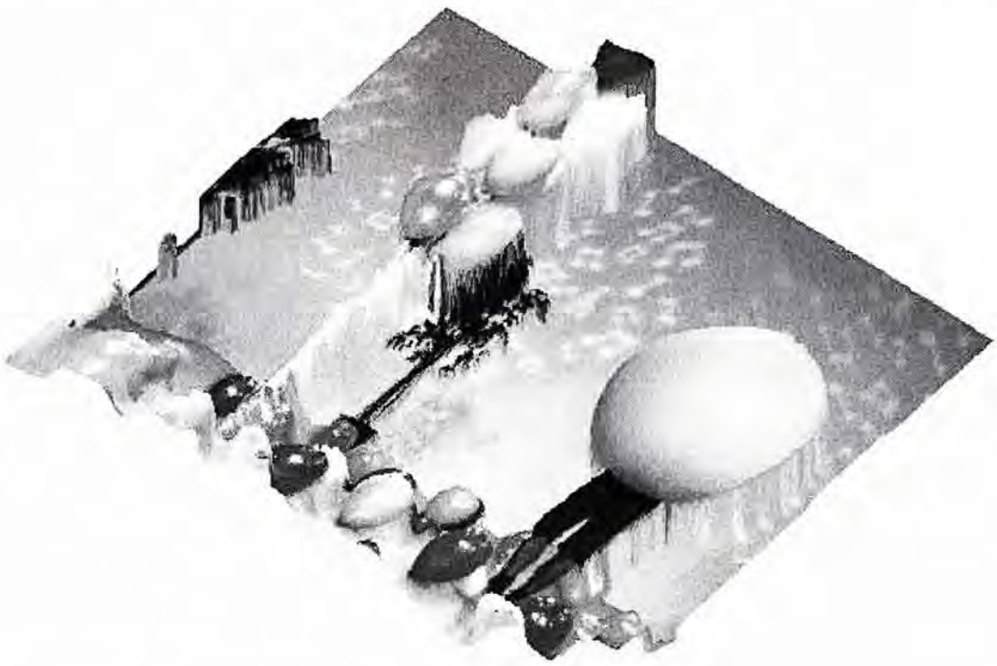


Figure 3.8: New-view synthesize using our depth map in Fig. 3.9.

3.4 Results

We show in this section several frames of two sequences to illustrate the improvement in depth estimate consistency and accuracy. Initial depths, final depths, and intermediate results are shown for comparison.

Fig. 3.9 and Fig. 3.10 show two sets of intermediate and final results output from our method. The first row shows the input images; the second row shows the initial depth maps that are not very consistent. The third row contains results of the trajectory-based edge profiles, faithfully enhancing boundaries. The forth row shows our depth maps after temporal refinement to improve the consistency. The fifth row contains the final depth results after sub-pixel continuous refinement. Even the thin edges are preserved well. Deforming objects are also satisfyingly handled. More frames of the inital and final depth maps are shown in Figs. 3.11, 3.12, 3.13 and 3.14.

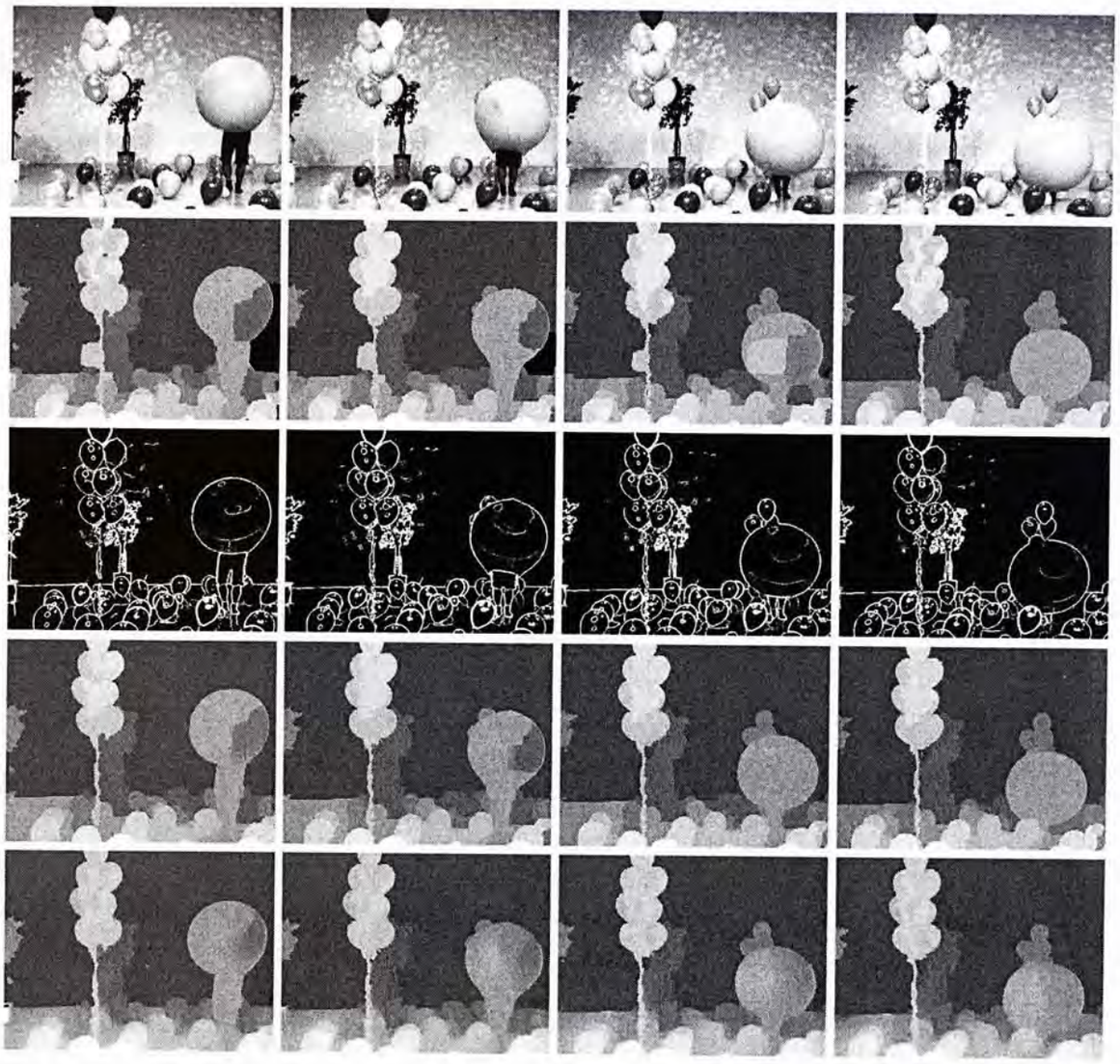


Figure 3.9: The balloon sequence results. The 1^{st} row: input images; the 2^{nd} row: initial depth maps; the 3^{rd} row: trajectory-based edge profiles; the 4^{th} row: depth maps after temporal refinement; the 5^{th} row: depth maps after sub-pixel continuous refinement.

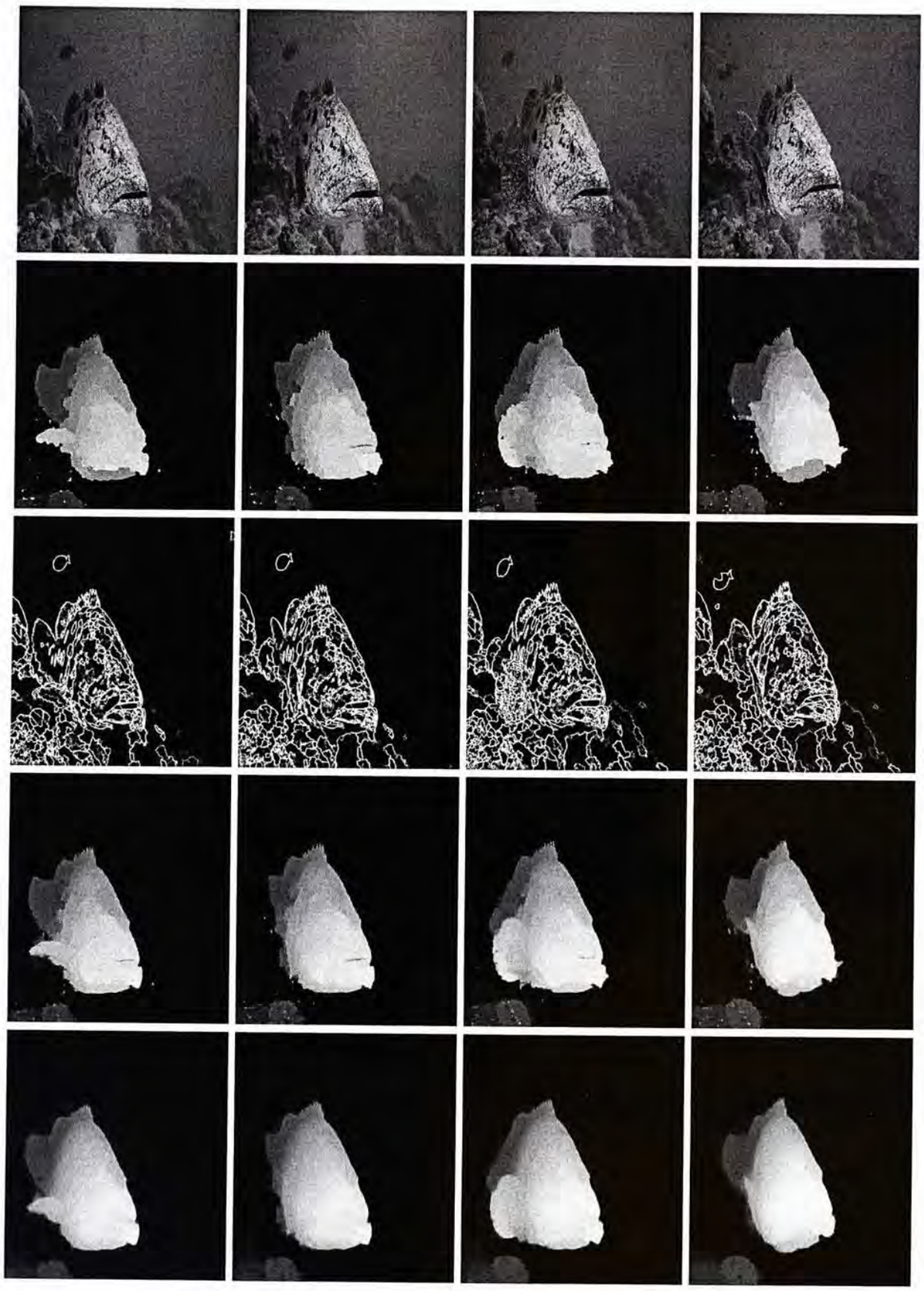


Figure 3.10: The fish sequence results. The 1^{st} row: input images; the 2^{nd} row: initial depth maps; the 3^{rd} row: trajectory-based edge profiles; the 4^{th} row: depth maps after temporal refinement; the 5^{th} row: depth maps after sub-pixel continuous refinement.

To demonstrate the accuracy of the depth, we synthesis a new view based on the computed depth maps. Note that continuous variation of depth is also preserved together with discontinuous object boundaries, thanks to the effective regularization, multi-frame profile construction, and robust optimization.

□ **End of chapter.**

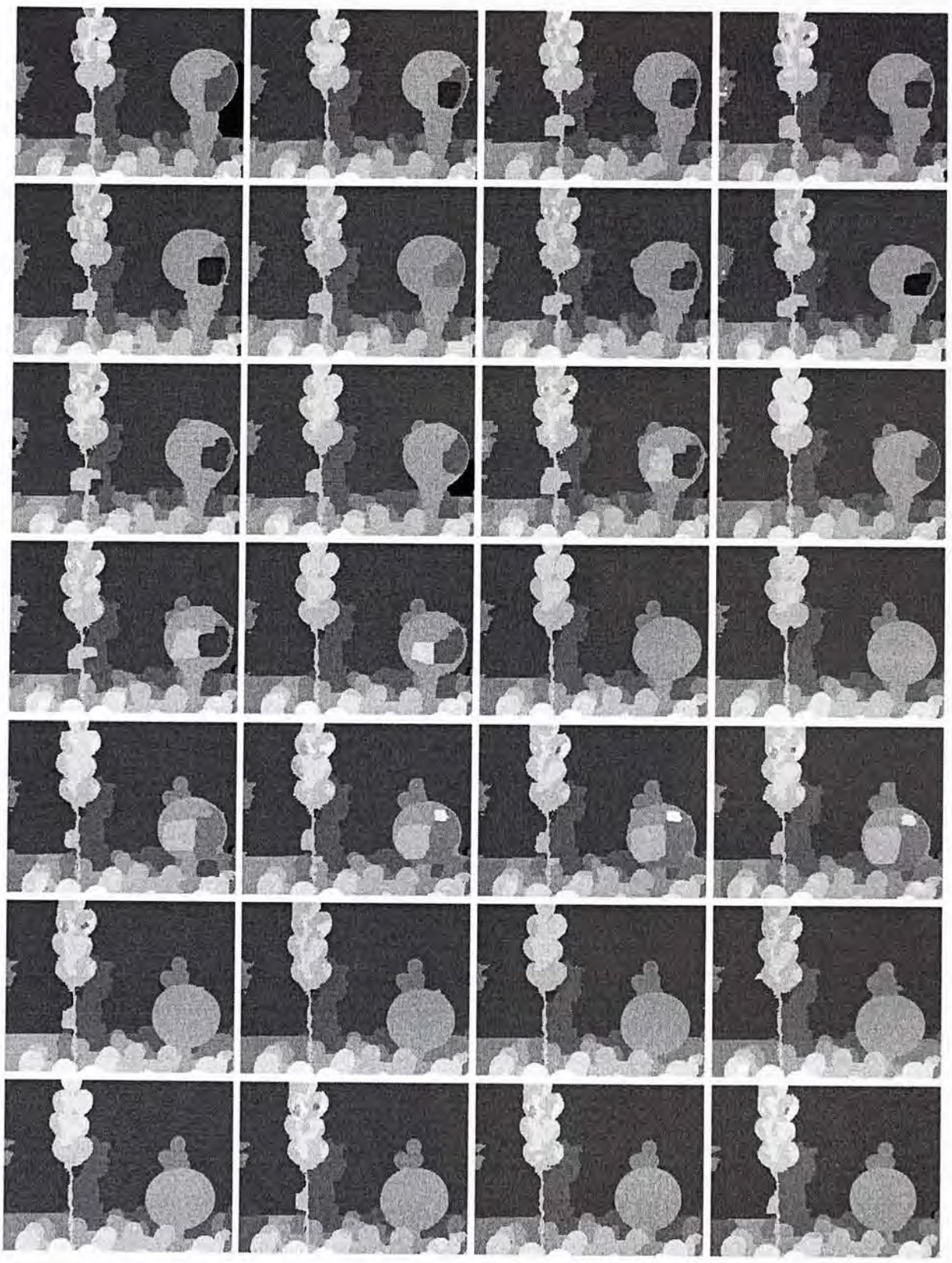


Figure 3.11: Frames 1 – 28 of initial depth maps of the balloon sequence.

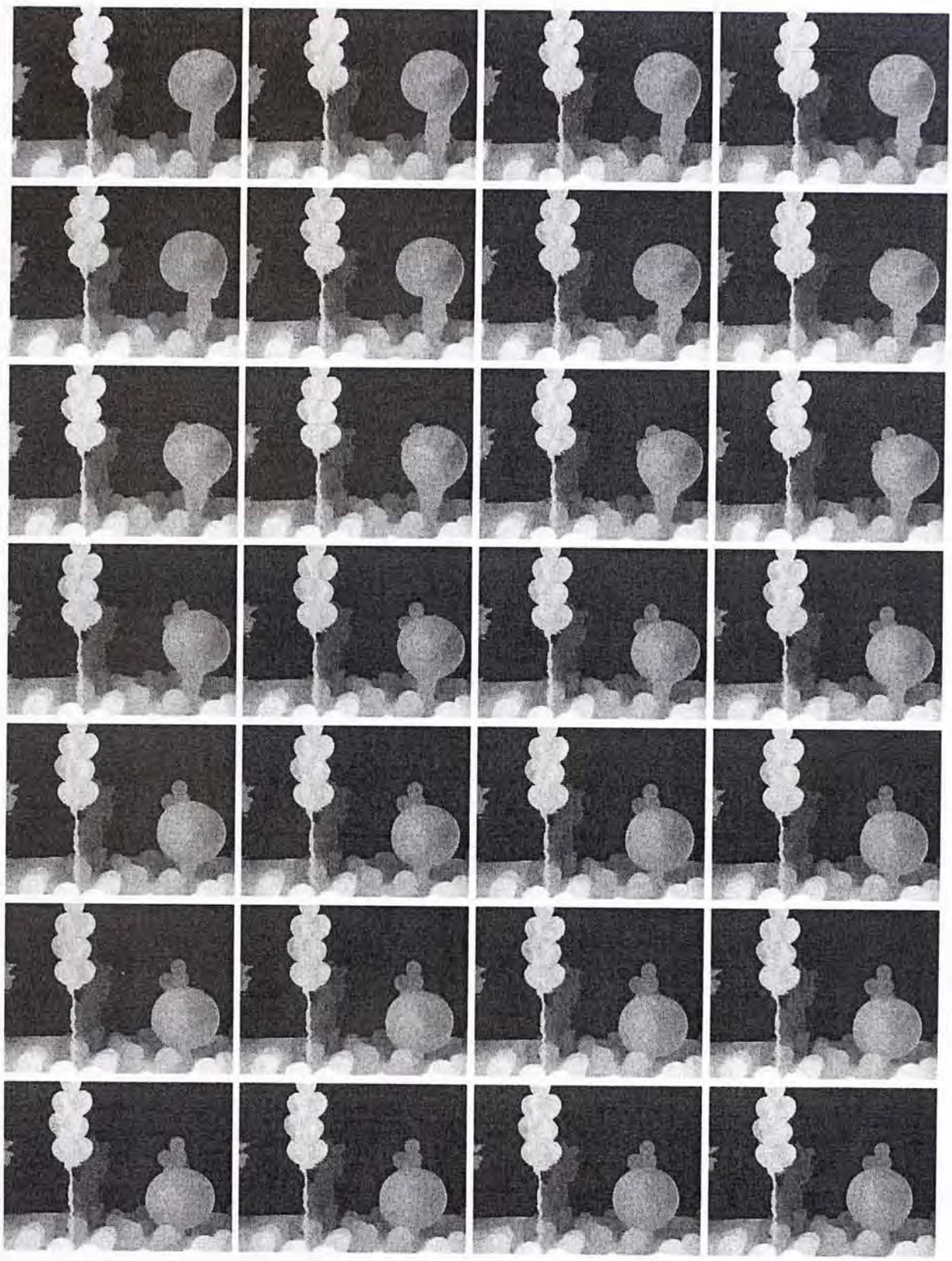


Figure 3.12: Frames 1 – 28 of final depth maps of the balloon sequence. The flickering artifacts are greatly reduced. The object boundaries are much improved.

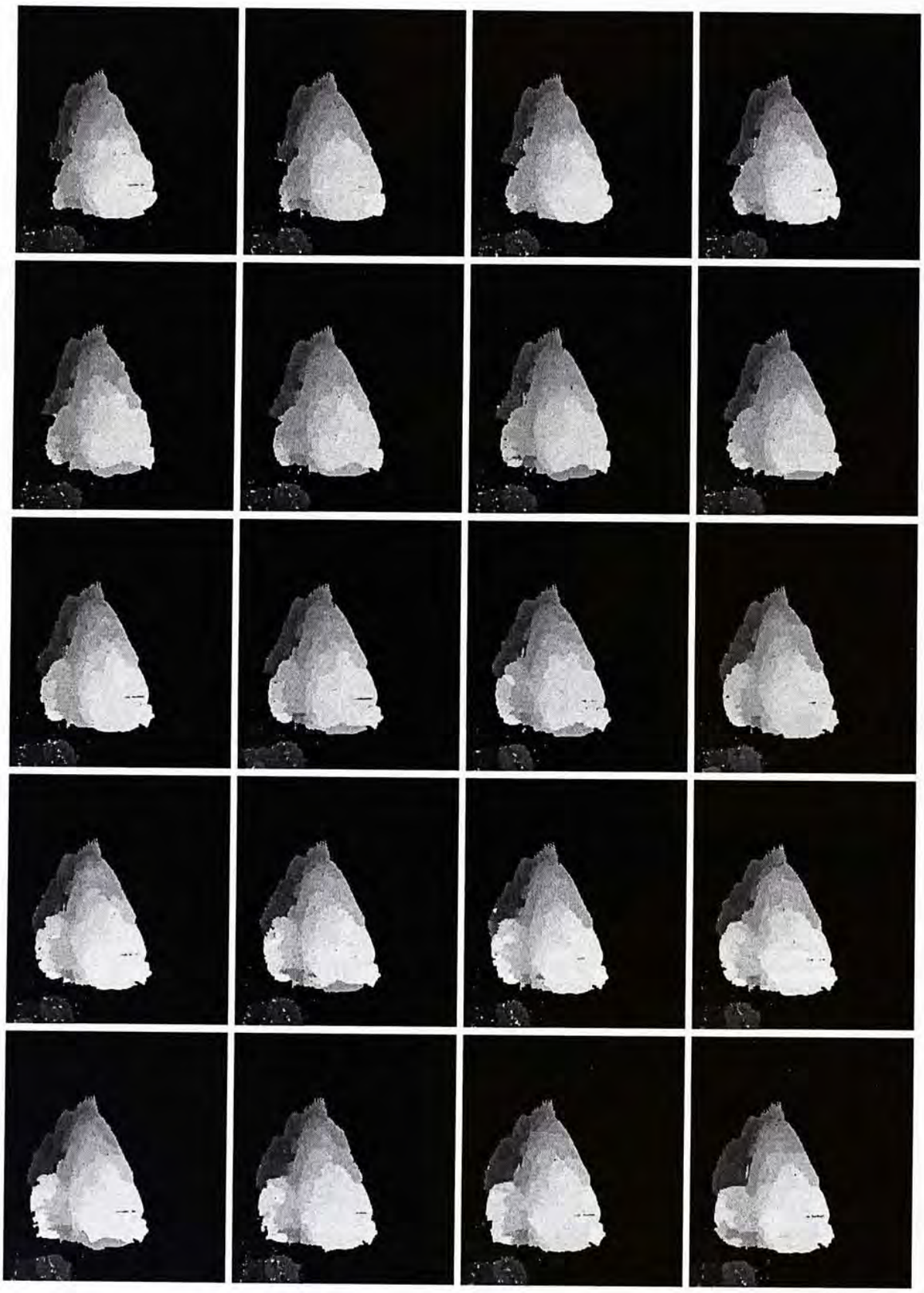


Figure 3.13: Frames 41 – 60 of initial depth maps of the fish sequence.

Figure 3.14: Frames 41 – 60 of final depth maps of the fish sequence.

Chapter 4

Conclusion

In this thesis, we generalized the depth estimation process in the traditional photometric stereo problem and in the traditional stereo matching problem.

For the proposed PS framework, we have presented a novel approach to mitigate the data capturing restriction for photometric stereo. By considering environment lighting, we regard both direct and indirect illumination as the same optical phenomenon, from which a simple but very effective framework was derived. Our method uses a mirror sphere to calibrate environment lighting, making complex indirect illumination no longer decisively harmful. Different environments were tested – from dark-room to outdoor scenes. In all of them, decent surface can be produced.

We analyzed the Lambertian case in relaxing the well-known dark-room constraint and believe our framework can be naturally extended to other analytical shading models, for example the Ward's model.

In summary of the stereo matching framework, the major novelty is three-fold. First, the motion trajectory construction can find reliable estimates consecutively and break links when occlusion consistently arises in multiple frames. Occasional noise in one or two frames, on the contrary, can be

robustly skipped. Second, the novel edge occurrence maps are constructed incorporating structural information from multiple frames. The voting-like average scheme greatly suppresses errors that cannot be consistent in many frames and enhance correct estimates. Third, we propose the anisotropic smoothing scheme to provide proper regularization for all pixels. Possible edges are only with isotropic directional smoothness penalty while flat regions are enforced smoothness in all directions. In Chapter 3.4, we show that these contributions make binocular depth estimation in long sequences temporally consistent, robust to noise and other visual artifacts.

□ **End of chapter.**

Bibliography

[1] A. Hosni, M. Bleyer, M. G., and Rheman, C. Local stereo matching using geodesic support weights. In *ICIP* (2009), pp. 2093–2096.

[2] Agrawal, A., and Raskar, R. What is the range of surface reconstructions from a gradient field? In *ECCV* (2006).

[3] Alldrin, N., Zickler, T., and Kriegman, D. Photometric stereo with non-parametric and spatially-varying reflectance. In *CVPR* (2008).

[4] Álvarez, L., Deriche, R., Papadopoulo, T., and Sánchez, J. Symmetrical dense optical flow estimation with occlusions detection. *International Journal of Computer Vision 75*, 3 (2007), 371–385.

[5] Basri, R., Jacobs, D., and Kemelmacher, I. Photometric stereo with general, unknown lighting. *Int. J. Comput. Vision 72*, 3 (2007), 239–257.

[6] Bobick, A. F., and Intille, S. S. Large occlusion stereo. In *IJCV* (1999), vol. 33, pp. 181–200.

[7] Boykov, Y., Veksler, O., and Zabih, R. Fast approximate energy minimization via graph cuts. *IEEE Trans. PAMI 23*, 11 (2001), 1222–1239.

[8] Brox, T., Bruhn, A., Papenberg, N., and Weickert, J. High accuracy optical flow estimation based on a theory for warping. In *ECCV (4)* (2004), pp. 25–36.

[9] Bruhn, A., and Weickert, J. Towards ultimate motion estimation: Combining highest accuracy with real-time performance. In *ICCV* (2005), pp. 749–755.

[10] Bruhn, A., Weickert, J., and Schnörr, C. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *IJCV 61*, 3 (2005), 211–231.

[11] Chung, H.-S., and Jia, J. Efficient photometric stereo on glossy surfaces with wide specular lobes. *CVPR* (2008).

[12] Coleman-Jr., E., and Jain, R. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Computer Graphics and Image Processing 18*, 4 (1982), 309–328.

[13] Debevec, P. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH* (1998), pp. 189–198.

[14] Esteban, C. H., Vogiatzis, G., and Cipolla, R. Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30* (2008), 548–554.

[15] Felzenszwalb, P. F., and Huttenlocher, D. P. Efficient belief propagation for early vision. 41–54.

[16] GOLDMAN, D. B., CURLESS, B., HERTZMANN, A., AND SEITZ, S. M. Shape and spatially-varying brdfs from photometric stereo. In *ICCV05* (2005), pp. 341–348.

[17] HERNANDEZ, C., VOGIATZIS, G., BROSTOW, G. J., STENGER, B., AND CIPOLLA, R. Non-rigid photometric stereo with colored lights. *ICCV07* (2007), 1–8.

[18] HERTZMANN, A., AND SEITZ, S. M. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 8 (2005), 1254–1264.

[19] HIGO, T., MATSUSHITA, Y., AND IKEUCHI, K. Consensus photometric stereo. In *CVPR* (2010), pp. 1157 - 1164.

[20] HOLROYD, M., LAWRENCE, J., HUMPHREYS, G., AND ZICKLER, T. A photometric approach for estimating normals and tangents. *ACM Trans. Graph. 27*, 5 (2008).

[21] HUGUET, F., AND DEVERNAY, F. A variational method for scene flow estimation from stereo sequences. In *ICCV* (2007), pp. 1–7.

[22] IRANI, M. Multi-frame correspondence estimation using subspace constraints. *International Journal of Computer Vision 48*, 3 (2002), 173–194.

[23] KOLMOGOROV, V., AND ZABIH, R. What energy functions can be minimized via graph cuts? *IEEE Trans. Pattern Anal. Mach. Intell. 26*, 2 (2004), 147–159.

[24] LIM, J., HO, J., YANG, M., AND KRIEGMAN, D. Passive photometric stereo from motion. In *ICCV05* (October 2005).

[25] MIN, D. B., AND SOHN, K. Edge-preserving simultaneous joint motion-disparity estimation. In *ICPR (2)* (2006), pp. 74–77.

[26] NAYAR, S., IKEUCHI, K., AND KANADE, T. Determining shape and reflectance of hybrid surfaces by photometric sampling. *IEEE Trans. Robo. Auto. 6*, 4 (1990), 418–431.

[27] NEHAB, D., RUSINKIEWICZ, S., DAVIS, J., AND RAMAMOORTHI, R. Efficiently combining positions and normals for precise 3D geometry. *ACM SIGGRAPH 2005 24*, 3 (Aug. 2005).

[28] NICODEMUS, F. E. Directional reflectance and emissivity of an opaque surface. *Applied Optics 4*, 7 (1965), 767–775.

[29] PATRAS, I., ALVERTOS, N., AND TZIRITAS, G. Joint disparity and motion field estimation in stereoscopic image sequences. In *International Conference on Pattern Recognition* (1996), vol. 1, pp. 359–363.

[30] PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., AND FLANNERY, B. P. *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press. New York. NY. USA. 1992.

[31] SAND, P., AND TELLER, S. J. Particle video: Long-range motion estimation using point trajectories. *International Journal of Computer Vision 80*, 1 (2008), 72 91.

[32] SATO, Y., AND IKEUCHI, K. Reflectance analysis under solar illumination. Tech. rep., Pittsburgh, USA, 1994.

[33] SHEN, L., AND TAN, P. Photometric stereo and weather estimation using internet images. *CVPR09* (2009), 1850–1857.

[34] SHI, B., MATSUSHITA, Y., WEI, Y., XU, C., AND TAN, P. Self-calibrating photometric stereo. In *CVPR* (2010), pp. 1118 - 1125.

[35] SOLOMON, F., AND IKEUCHI, K. Extracting the shape and roughness of specular lobe objects using four light photometric stereo. *IEEE Trans. PAMI 18* (1996), 449–454.

[36] TAN, P., LIN, S., AND QUAN, L. Subpixel photometric stereo. *IEEE Trans. PAMI 30*, 8 (2008), 1460–1471.

[37] VALGAERTS, L., BRUHN, A., ZIMMER, H., WEICKERT, J., STOLL, C., AND THEOBALT, C. Joint estimation of motion, structure and geometry from stereo sequences. In *ECCV (4)* (2010), pp. 568–581.

[38] VEDULA, S., BAKER, S., RANDER, P., COLLINS, R. T., AND KANADE, T. Three-dimensional scene flow. *IEEE Trans. Pattern Anal. Mach. Intell. 27*, 3 (2005), 475–480.

[39] WEDEL, A., RABE, C., VAUDREY, T., BROX, T., FRANKE, U., AND CREMERS, D. Efficient dense scene flow from sparse or dense stereo data. In *ECCV (1)* (2008), pp. 739–751.

[40] WOODHAM, R. Photometric method for determining surface orientation from multiple images. *Optical Eng. 19*, 1 (1980), 139–144.

[41] WU, T.-P., TANG, K.-L., TANG, C.-K., AND WONG, T.-T. Dense photometric stereo: A markov random field approach. *IEEE Trans. Pattern Anal. Mach. Intell. 28*, 11 (2006), 1830–1846.

[42] XU, L., JIA, J., AND MATSUSHITA, Y. Motion detail preserving optical flow estimation. In *CVPR* (2010), pp. 1293–1300.

[43] Zhang, Y., and Kambhamettu, C. On 3d scene flow and structure estimation. In *CVPR (2)* (2001), pp. 778–785.

[44] Zhang, Z., and Faugeras, O. D. Estimation of displacements from two 3-d frames obtained from stereo. *IEEE Trans. Pattern Anal. Mach. Intell. 14*, 12 (1992), 1141–1156.

[45] Zimmer, H., Bruhn, A., Weickert, J., Levi Valgaerts and, Agustín Salgado, B. R., and Seidel, H.-P. Complementary optic flow. In *EMMCVPR* (2009).